INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




AI-DRIVEN EMOTION RECOGNITION FOR MENTAL HEALTH DIAGNOSES: Assessing Mental Health through Emotional State Evaluation

PEDRO MIGUEL PERES PRETO
Outubro de 2025

P.PORTO

AI-DRIVEN EMOTION RECOGNITION FOR MENTAL HEALTH DIAGNOSES

Assessing Mental Health through Emotional State Evaluation

Pedro Miguel Peres Preto

Student No.: 1190963

Dissertation for the Degree of

Master in Artificial Intelligence Engineering

Supervisor: Luís Manuel Silva Conceição

Co-Supervisor: Ana Maria Neves Almeida Baptista Figueiredo

Jury:

President:

António Constantino Lopes Martins, Associate Professor, Institute of Engineering, Polytechnic of Porto

Vowels:

Jorge Fernandes Rodrigues Bernardino, Associate Professor, Institute of Engineering, Polytechnic of Coimbra

Luís Manuel Silva Conceição, Assistant Professor, Institute of Engineering, Polytechnic of Porto

Porto, September 2025

Dedicatory

To my family for their foundation, and to myself for building upon it.

Abstract

Mental health conditions remain a concerning challenge across the globe, requiring timely and reliable approaches to correctly make accurate diagnoses and effective interventions. Traditional assessment methods often rely on subjective self-reports and clinical interviews, which may not always capture the full spectrum of an individual's emotional state. In this context, computational techniques for emotion analysis provide a complementary perspective by identifying patterns in facial expressions, speech, and language.

This dissertation evaluates the potential of multimodal emotional state analysis and its contribution to mental health assessment, through the development of a computational application. A systematic review was conducted to evaluate existing methodologies and highlight their strengths, limitations, and applicability in clinical contexts. Building on this review, the present work explores an integration of visual, vocal, textual patterns, assessing the contribution of their combined capacity to improve the consistency and depth of emotional interpretation.

An analysis centered on methodological design was conducted by applying techniques such as preprocessing, fine-tuning, and data augmentation on the datasets to enhance the model's capacity. Ethical and security considerations were also incorporated to strengthen system robustness and ensure responsible deployment in the market.

The proposed solution consists of an artificial intelligence based multimodal system that integrates the analysis of emotions present in facial expressions, voice, and text patterns to provide a comprehensive assessment of the user's emotional state. The application's modular architecture enables real-time processing and the generation of clinical reports.

The experimental validation of the system revealed promising results across several DSM-5 domains, the clinical reference manual that defines diagnostic criteria for mental disorders cases. High F1-scores were recorded in domains such as Anger (0.84) and Personality Functioning (0.87), while more subtle domains, such as Dissociation (0.43) and Repetitive Behaviors (0.52), revealed more modest performance. The overall analysis resulted in an observed agreement level of 71.9% and a Cohen's Kappa of 0.42, indicating moderate agreement with the DSM-5.

The findings underline the promise of computational emotion analysis as a supplementary tool for mental health professionals, while also emphasizing the importance of critical evaluation of its limitations and careful integration into clinical practice.

Keywords: mental health, AI-driven multimodal analysis, clinical decision support, facial expression analysis, speech processing, sentiment analysis.

Resumo

As condições de saúde mental continuam a representar um desafio preocupante em todo o mundo, exigindo abordagens rápidas e fiáveis que permitam diagnósticos precisos e intervenções eficazes. Os métodos de avaliação tradicionais baseiam-se, frequentemente, em autorrelatos subjetivos e entrevistas clínicas, que nem sempre permitem uma perceção completa do estado emocional de um indivíduo. Neste contexto, as técnicas computacionais de análise de emoções oferecem uma perspetiva complementar ao identificar padrões em expressões faciais, no tom de voz e no discurso.

Esta dissertação avalia o potencial da análise multimodal de estados emocionais e o seu contributo para a avaliação da saúde mental, através da criação de um sistema computacional. Foi realizada uma revisão sistemática, para avaliar as metodologias existentes e destacar os seus pontos fortes, limitações e aplicabilidade em contextos clínicos. Com base nesta revisão, o presente trabalho, explora a integração de padrões visuais, vocais e textuais, analisando o contributo da sua combinação para uma interpretação emocional mais concisa e aprofundada.

Foi realizada uma análise centrada no *design* metodológico, aplicando técnicas como pré-processamento, *fine-tuning* e *data augmentation* nos *datasets*, de forma a potenciar a capacidade dos modelos. Foram igualmente consideradas dimensões éticas e de segurança, para reforçar a robustez do sistema e assegurar um lançamento responsável no mercado.

A solução proposta consiste num sistema multimodal baseado em inteligência artificial que integra a análise de emoções presentes em expressões faciais, na voz e no texto, de modo a fornecer uma avaliação abrangente do estado emocional do paciente. Ademais, a arquitetura modular da aplicação permite o processamento e a geração de relatórios clínicos em tempo real.

A validação experimental do sistema revelou resultados promissores em diversos domínios do DSM-5 – manual de referência clínica que define critérios de diagnóstico para casos de transtornos mentais. Registaram-se *F1-scores* elevados em domínios como a Raiva (0.84), e Funcionamento da Personalidade (0.87), enquanto domínios mais subtis, como Dissociação (0.43) e Comportamentos Repetitivos (0.52) apresentaram desempenhos mais modestos. A análise global resultou num nível de concordância observada de 71.9% e num Kappa de Cohen de 0.42, o que indica uma concordância moderada com o DSM-5.

Os resultados reforçam o potencial da análise computacional de emoções enquanto ferramenta suplementar para profissionais de saúde mental, salientando, contudo, a necessidade de uma avaliação crítica das suas limitações e da integração cuidada na prática clínica.

Palavras-Chave: saúde mental, análise multimodal baseada em inteligência artificial, apoio à decisão clínica, análise de expressões faciais, processamento de fala, análise de sentimentos.

Acknowledgement

To commence, I would like to thank my family. To my father for teaching me through his example the profound importance of hard work and perseverance. To my mother, for being my safe port and for always being there in the hours of greatest need. To my brother, for... well, for providing an excellent lesson in patience and for demonstrating a level of unshakable self-confidence that I deeply admire.

To my girlfriend for helping me through the obstacles of life and for always giving me kind words during the toughest days.

To my friends, who believe and root for me unconditionally. Your faith is a privilege and a gift that I carry every day.

To my thesis supervisor, Luis Conceição, I wish to express my special gratitude for his guidance and patience throughout this journey. Finally, to my co- supervisor, Ana Almeida, I would like to outline my appreciation for her essential help and for being a pillar of strength throughout this process.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
AI Act	Artificial Intelligence Act
AUC	Area Under the Curve (ROC curve metric)
BoME	Body Motor Elements
DPIA	Data Protection Impact Assessment
STFT	Chroma Short-Time Fourier Transform
CNN	Convolutional Neural Network
DSR	Design Science Research
ER	Emotion Recognition
ERS	Emotion Recognition System
GDPR	General Data Protection Regulation
KNN	k-Nearest Neighbors
LMA	Laban Movement Analysis
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficient
NLP	Natural Language Processing
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QCNN	Quantum Convolutional Neural Network
RMS	Root Mean Square
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
TSA	Textual Sentiment Analysis
VER	Voice Emotion Recognition
ZCR	Zero Crossing Rate

1 Introduction

This chapter presents the context of this study and justifies the relevance of the proposed system. Additionally, it introduces the focus of the research along with its goals. Finally, it exposes the adopted research methodology and the document structure.

1.1 Context

Mental health has become an increasingly global issue, which impacts individuals all around the world. The World Health Organization study affirms that more than 264 million people around the globe deal with depression, being worldwide, one of the top reasons for disability [1]. Looking back to 2010, the economic impact from mental disorders was around 2.5$ trillion, value that is expected to double by 2030 [2]. The Corona Virus Disease (COVID-19) made things even tougher, causing a big increase in mental health problems. For example, anxiety levels rose by about 25% to 30% from 2010 to 2020 [1]. The lockdown during this pandemic highly contributed to the increase in mental disorders, as individuals faced social isolation, economic difficulties, and difficult access to health services. Furthermore, even solutions like COVID-19 vaccine raised safety concerns due misinformation and conspiracy theories about its side effects [3].

Mental illness often manifests as changes in mood, posture and behavior towards others. Despite that, the subjective criteria of traditional diagnostic approaches, like patient interviews and self-reported experiences, often fail to provide precise assessments ahead of time. In only 2019, eight hundred thousand suicides were recorded globally. It was stated that the core contributor to all these documented deaths were health illnesses. Additionally, in India, a survey revealed that 1 in 20 individuals suffers from depression, which contributes for the rise of suicide, mainly among people with ages comprehended between 15 and 29 years old, and on a worldwide scale, suicide was confirmed as the second leading cause of death. This outcome could have been avoided by an earlier diagnosis [1].

1.2 Motivation

Given the widespread impact and cost of mental health disorders, we need new approaches that move past traditional methods. Early and more objective diagnoses could prevent tragic outcomes like suicide and improve patients' quality of life.

Emerging technologies and techniques based on Artificial Intelligence (AI) – deep learning, machine learning and computer vision – show real potential to transform mental healthcare. Emotion detection, sometimes, may go unnoticed in traditional medical-patient appointments and such information can be crucial for the psychologist to make a precise health evaluation of this patient. Among these innovative ways to support the mental health field, facial expression analysis stands as one of the most effective non-verbal methods for understanding emotional states. The continuous advancements in deep learning, including convolutional neural networks (CNNs) and attention mechanisms, made systems able to analyze facial expressions, a non-intrusive method, alongside contextual patterns, such as body posture, to achieve a higher accuracy in emotion recognition [4], [5]. At the same time, Natural Language Processing (NLP) and sentiment analysis provide complementary insights by examining speech tone, patterns, and vocabulary [6], [7].

1.3 Problem Statement

Mental health disorders are becoming, day by day, an increasing global issue. However, current diagnostic methods still rely heavily on subjective self-reports and clinical interviews. These methods lack objectivity, speed, and consistency. As a result, they may delay early interventions and lessen the effectiveness of mental healthcare.

This dissertation addresses the problem of the lack of objective, real-time systems that can help clinicians recognize and interpret patients' emotional states. This issue is especially important in telemedicine, where fewer face-to-face interactions increase the chances of misunderstandings and incomplete evaluations. To tackle these problems, this thesis aims to create an AI-powered multimodal system, leveraging a combination of machine and deep learning techniques, to improve diagnostic reliability and provide real-time alerts to healthcare professionals.

On the bottom line, the goal of the work is to improve mental health assessment, and ultimately, provide a more accurate, earlier, and effective patient-driven intervention.

1.4 Objectives

The primary goal of this research is to develop an AI-driven system that enhances mental health diagnostics and personalized therapeutic interventions by using AI-powered emotion recognition techniques. The proposed system aims to bridge the gap between technological advancements and real-world applications, providing healthcare professionals detailed reports

about their doctor-patient appointments and with real-time insights about patient emotional state. Therefore, the central research question of this study is:

"How can an AI-powered system leverage emotion recognition techniques to improve mental health diagnostics and therapeutic interventions?"

Hence, to answer the aforementioned question the following objectives will be pursued:

- Development of an emotion recognition system that integrates a multimodal emotion recognition approach – utilizes different contextual information's to improve the reliability of emotion detection during the appointments and medical sessions.
- Implementation of mechanisms to generate professional reports, based on the information's collected during the sessions, providing important insights into patient emotional patterns.
- Design and integration of a module capable of generating personalized therapeutic suggestions and clinical recommendations by combining emotion recognition data with relevant clinical information.
- Build a real-time notification procedure to notify mental healthcare professionals about possible emotional changes detected in the patients.
- Integration of the solution as a plug-in or add-on, for Zoom or Teams, to allow real-time emotion detection during online appointments, enhancing the usability of the system for mental health professionals.

The main objective of this study and question is to contribute to the evolution of mental health care systems, combining innovative emotion recognition techniques, following a multimodal approach, to offer efficient and personalized mental health support.

1.5 Contributions

This dissertation aims to contribute to mental health diagnostics by using AI-driven emotion recognition. The key contributions can be briefed as follows:

1. **Enhanced Diagnosis Precision**: By using literature emotion recognition methodologies, the system aims to improve the accuracy and precision of emotion recognition, which is directly related to the improvement of mental health diagnosis's reliability.
2. **Real-Time Monitoring:** Integrating real-time feedback on the patient's emotions helps health professionals in tracking patients' mental developments and adjust interventions dynamically.
3. **Personalized Reports:** Developing a module that generates tailored diagnostic reports and therapeutic recommendations based on the doctor-patient session.
4. **Accessibility and Scalability:** The goal is to design a system that can function efficiently in diverse environments, including remote environments by using the developed application or even plugins for Zoom and Microsoft Teams.

This work aims to lay a foundation for a transformative approach on mental health diagnostics, leveraging cutting-edge technologies enhancing accessibility, accuracy and personalized mental care.

1.6 Research Methodology

This study will follow the Design Science Research (DSR) methodology as its guiding framework. DSR is a research paradigm that addresses the development and evaluation of innovative artifacts designed to tackle real-world problems [8], [9]. Its systematic and iterative nature makes it ideal for facing challenges and practical problems [9].

In the high-demanding mental health environment, where early and precise diagnosis can distinguish between life and death, the need for innovation is inevitable. DRS's provides a structured solution on artifact creation and enhancement [8] which aligns with the study goals – solve real-world practical challenges and contribute to theoretical advancements.

Accordingly, to Preffers et al., the DRS methodology (cf. Figure 1) proposes a six-research-step [9]:

1. **Identify the Problem & Motivation**: The first step is to define the research problem, which in this case involves need for a tool design to assist mental health professionals in diagnosing mental health diseases and providing therapeutic interventions. For this reason, a systematic review using the PRISMA methodology [10] will be conducted to explore the main questions within this subject.
2. **Define Objectives of a Solution**: Once the problem is identified, the following step is to structure objectives for a solution. In this study, the primary goal is to create an emotion recognition system that can offer professionals innovative and accurate means to evaluate mental health states.
3. **Design & Development**: This stage requires the design and development of the solution. By leveraging cutting-edge technology, the solution conceptualization aims to give precise insights on the patient's mental health and suggest personalized interventions.
4. **Demonstration**: The developed artifact is tested using a real-world scenario to validate its effectiveness and performance in solving the identified problem.
5. **Evaluation**: The system assessment metrics such as accuracy, recall, precision and F1 score were used to prove its effectiveness.
6. **Communication**: The concluding step comprises the study findings and contributions dissemination. This research was structured as a master's thesis in Artificial Intelligence, providing innovative insights into mental health and emotion recognition fields. Additionally, an article will be written and submitted for publication in relevant academic journals that focus on the role of emotion recognition on mental health diagnoses, ensuring that the findings reach more professionals in the field.

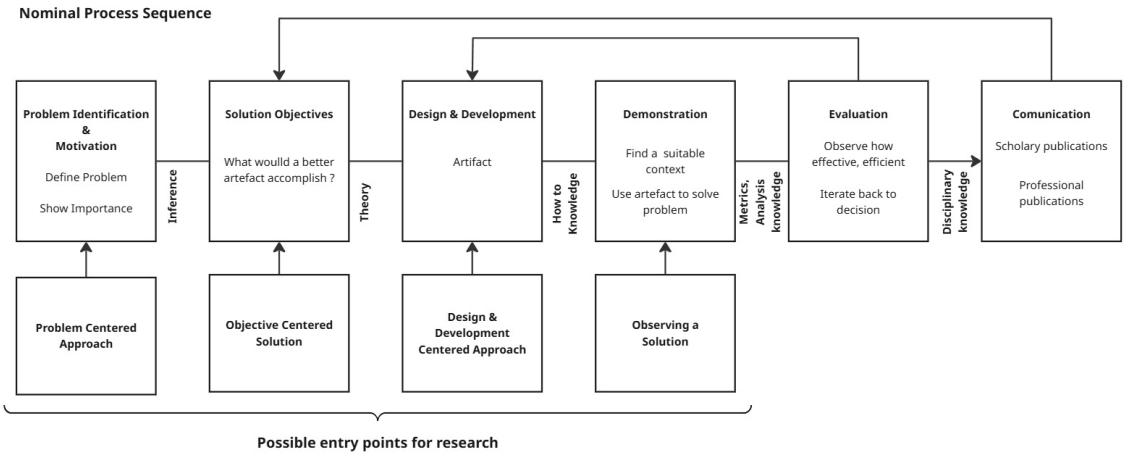

Figure 1 – DSR Methodology [11]

1.7 Document Structure

This study is structured into five chapters, each addressing a specific aspect of the research. The first chapter provides the background for the study and gives grounds for the proposed emotion recognition system. It also highlights the contributions made and explains the research methodology utilized.

The second chapter introduces the state of the art regarding emotion recognition on mental health analysis. It examines the key aspects of the technology and conducts a systematic review with the purpose of answering the primary research question of the study.

The third chapter explores the proposed model, describing the methods, tools and evaluation approaches used, while also presenting the concerns related to data protection, security and ethical considerations.

The fourth chapter focuses on the implementation, analysis, and discussion of the results. It includes a comparative analysis between metrics and the proposed model results.

The fifth and last chapter concludes the study by summarizing the main findings thar were drawn from the development and results analysis, highlighting the limitations encountered and suggesting future work proposals.

2 State of Art

The following chapter is divided into two subsections. The first section is a theoretical contextualization and highlights the procedures and methodologies employed in the current state-of-the-art emotion analysis systems. Additionally, it presents the main approaches and techniques applied to its development, as well as the key problems, limitations, and metrics used to evaluate them.

In the second section, a systematic review is conducted utilizing the PRISMA methodology. This review answers the proposed research questions (cf. 2.2.2 Research Questions), providing guidance and innovative knowledge for this study's emotion recognition system.

2.1 Emotional Recognition Systems

Emotional Recognition Systems (ERS) represent advanced technological techniques designed to identify and understand human emotions through the interpretation of specific data types (e.g., facial expressions, voice and text, body movements, etc.), each one characteristic of a different emotion recognition methodology. These systems have found extensive use in many domains, including mental health care, education, and customer support, as they provide important insight into emotional states and can support adaptive communication and intervention strategies [12], [13].

These systems utilize cutting-edge AI techniques to analyze data from many sources. The adoption of these approaches is highly accurate and proves to be effective while processing complex datasets, allowing real-time emotion recognition [14]. Acting as an intermediary layer between expressing human emotion and receiving computational understanding, these systems form the foundation for real-world customized applications interactions [14], [15].

2.1.1 Emotional Recognition Systems Methodologies

Emotional recognition systems employ diverse methodologies to capture and analyze human emotions. These approaches can be broadly categorized based on the type of input data that they analyze. In literature, four main categories stand out: facial expression analyses, speech and vocal emotion recognition, textual sentiment analysis, and psychological signal monitoring. Every single of these methodologies provides distinctive information into emotional states. Facial expressions give non-verbal signs, speech reveals acoustic and paralinguistic features, textual data captures semantic and syntactic aspects of language, and psychological signals render biometric indications of affective responses.

1. Facial Expression Analysis

Facial expression analysis plays a vital role in emotional recognition systems, especially in mental health settings, since facial expressions serve as dependable indicators of emotional states and underlying psychological issues (e.g., depression, anxiety, etc.) [16], [17], [18]. The universal nature of emotions conveyed through facial muscle movements —happiness, sadness, anger, and fear— makes facial recognition a powerful tool for assessing non-verbal communication [19], [20].

CNNs development was a major change to the field of facial expression recognition. Newer deep learning models, such as AlexNet and ResNet, have the capability to automatically create features by producing very reliable results, thus they surpass the older methods like Support Vector Machines (SVMs). In the case of AlexNet, it is proficient at extracting features from its fully connected layers, and when combined with classifiers like Linear Discriminant Analysis, it has shown strong performance on benchmark datasets including FER2013, AffectNet, and KDEF [21], [22]. Uniformly, ResNet's residual layers enhance the learning process for intricate emotional features, especially when dealing with large datasets [23].

Transferring learning has become a powerful technique for enhancing the performance of a model. Stronger models and good results have been achieved in real-world scenarios irrespective of the change in the lighting, posing and quality of the image by fine-tuning the pre-trained networks with datasets like FER2013, CFEE and JAFFE [24], [25].

Hybrid approaches that integrate static image analysis with temporal dynamics of video streams are on the rise. For example, models that emphasize the dynamism of emotions over time can depict the intricate nature of emotional states. Such systems have shown good performance in real-world clinical and educational scenarios as well as controlled settings [22], [24].

Quantum innovations have added more features to CNNs. Quantum Convolutional Neural Networks (QCNNs) are efficient whereby they cut costs by reducing training duration but at the same time increasing precision. Notably, these models outperformed others in recognizing multi-dimensional complex emotions in various subjects and conditions [26].

Multimodal approaches that unify facial analysis with body movement, voice and context are gradually gaining popularity and creating a trend. Other analyses combining facial feature

recognition with cross-culture Laban Movement Analysis of posture and gestures prove to be most valuable in appreciation of feelings. This multimodal integration of emotion recognition techniques into a singular system, ensures reliability and accuracy since it considers various dynamic features [25], [27].

2. Speech and Vocal Emotion Recognition

Speech and vocal emotion recognition (VER) are key components of modern emotional recognition systems since they utilize acoustic and linguistic features of speech to classify emotions. VER is especially important in mental health diagnostics, human-computer interaction and adaptive learning environments where vocal data is an easy and non-intrusive way to obtain and thereby infer emotional state [28], [29].

Essentially, VER systems extract and analyze many vocal features to determine emotions. Prosodic features like pitch, intensity and rate provide hints on the speaker's emotional tone. Spectral features like Mel-Frequency Cepstral Coefficients and Linear Predictive Cepstral Coefficients describe the timbre and resonance of speech. Furthermore, voice quality features like jitter, shimmer and Harmonics-to-Noise Ratio also reflect the physiological changes caused by emotional arousal [28], [29].

Throughout the years, the evolution of machine learning and deep learning caused a revolution in VER. Early approaches that made use of standard algorithms, like SVMs and k-Nearest Neighbors (KNN) revealed some limitations like struggles with generalization and often required hand-crafted features [29]. On the other hand, modern applications have overcome these setbacks by leveraging deep learning techniques such as CNN and Long Short-Term Memory (LSTM) networks, allowing the automated extraction of features and capturing temporal dependencies in speech [28], [30].

The integration of multimodal approaches has also increased the scope of VER systems. These systems can recognize emotional states better by incorporating vocal data, facial expressions, text and physiological signals. For example, in audio-visual emotion recognition methods, synchronization of facial expressions and speech enhances classification accuracy, particularly during clutter environments. Furthermore, decision-level fusion approaches, which aggregate predictions from many modalities, are invaluable in improving system reliability and precision [17], [21], [31].

3. Textual Sentiment Analysis

Textual Sentiment Analysis (TSA) has transformed into a crucial technique in understanding emotions and states of mind as portrayed in the text. The relationship of the written word to emotion, to the disorder diagnosis, and the treatment, in turn, determines the analysis value in the mental health domain [32].

TSA utilizes modern Natural Language Processing and machine learning methods in a bid to detect emotions from written content. One possible implementation is the application of deep learning techniques like Convolutional Neural Networks and Long Short-Term Memory

networks. With these models, emotions can be classified accurately due to their ability to detect syntactic and semantic features of text. For example, Bi-LSTM models refined with attention mechanisms have been able to find more complicated emotions in longer texts, given the many emotions involved with mental illness.

The datasets also significantly support the teaching of TSA systems. For instance, the DepressionEmotion dataset which contains mental health-related text with sobering annotations was instrumental in training multilabel emotion classification models. Multifaceted human annotation, along with training accomplished language models, also appears in the data because these sources use emotion detection models specifically for sadness, anger, and anxiety [33].

The use of integrations of word embeddings like GloVe or contextual embeddings such BER, have contributed towards improving the accuracy of sentiment analysis systems. The use of such embeddings enables the model to meaningfully utilize amorphous or intricately dependent sentiments as these embeddings store both local as well as global structural information. Multi-head attention mechanisms also improve these models by emphasizing critical portions of the text relevant to the emotional states [32], [34].

One major trend in TSA that has made a difference is the fusion of multimodal systems, in which text is combined with video or audio data. For instance, in mental health, diagnosing patients with depression or anxiety requires constant combining of textual sentiment with voice and facial expression, which enhances the context. Techniques such as decision-level fusion improve noise and data sparsity by combining outputs from multiple models whereas feature fusion combines the embeddings of text with prosodic and visual features [24].

4. Psychological Signals

Psychological signals such as electrooculogram, electromyogram, galvanic skin response, temperature, and heart rate variability are commonly utilized for emotion analysis due to their objectivity and low-power dissipation [35]. However, they usually require either wearable or intrusive procedures – neither of which would be practical during standard nor online psychologist-patient consultations.

It is worth mentioning that some non-contact approaches have been proposed in recent years. For example, infrared thermal cameras can estimate skin temperature, superficial blood flow variations, and stress-related stimulation without physical contact [36], [37]. Likewise, high-resolution remote eye trackers can capture gaze direction, blink rate, and pupil dilation without the need for headsets or contact lenses, and these features have been linked to emotional tone and arousal [38]. Similarly, radar and Wi-Fi based biosensing methods have been explored, which use radio signals to detect respiration and heartbeat without attaching sensors to the body [39].

Even with the above possibilities, they remain unsuitable for any form of psychological intervention. Generally, they rely on costly or specialized equipment that is not easily available on clinical practice, and they depend on controlled conditions such as stable lighting, fixed

positioning, or dedicated sensors. Also, such scenarios are not geared for telemedicine, as most of the devices rely on standard microphones and web cameras. Though these methods are contactless, these devices are still perceived as intrusive, and this could undermine the basic therapeutic contact of safety and trust. Therefore, the current conjecture on not making non-contact physiological sensing a viable alternative stand.

2.1.2 Metrics for Emotional Recognition Systems Evaluation

Measuring the emotion detection systems 'performance is a critical factor as it has a direct impact on trust and the reticence with which it is employed in any particular field. Several metrics are commonly used in the evaluation process and these metrics give many details about the advantages and disadvantages of the models.

2.1.2.1 Accuracy

Accuracy measures the proportion of correctly classified instances among the total predictions made by the model. It is defined as [23]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

TP stands for True Positives, TN represents True Negatives, FP represents False Positives, and FN Represent False Negatives. It is important that improved accuracy is achieved in areas where general dependability is of the utmost importance [23].

2.1.2.2 Precision

Precision evaluates the accuracy of the model in predicting true positives by calculating the percentage of true positives to all positive predictions made by the model. It is calculated as [23]:

$$Precision = \frac{TP}{TP + FN} \quad (2)$$

This is a critical metric when the cost of false positives is high [23].

2.1.2.3 Recall

Recall, often referred to as sensitivity, assesses how well the model can recognize all pertinent instances. It is represented as: It is expressed as [23]:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Recall plays a vital role in areas like mental health detection, where overlooking an emotional cue can have serious repercussions [23].

2.1.2.4 F1-score

F1-score offers a balance between precision and recall, making it particularly valuable when working with imbalanced class distributions. It is defined as the harmonic average of precision and recall [23]:

$$F1 = \frac{Precision \times Recal}{Precision + Recal} \quad (4)$$

An F1-score nearing 1 signifies an ideal equilibrium between precision and recall [23].

2.1.2.5 Other Metrics

The **ROC curve** demonstrates the balance between sensitivity and specificity at different threshold levels, and the area under the curve (AUC) measures the model's overall performance [23]. Furthermore, **Standard deviation** assesses how consistent the model's performance is across various datasets and situations, with lower values suggesting improved generalization [23].

2.1.3 Emotional Recognition Systems Challenges

Every system can face challenges and emotion analysis applications are no different. The primary challenges are listed below and encompass technical, ethical, and practical aspects:

Data-Related Problems: One of the significant bottlenecks is the need to source and create datasets that can be used in training and validating the various models built for emotional recognition. The range of emotions arising from cultures and individual differences complicates the creation of different datasets which account for all the anticipated scenarios. For example, datasets like AffectNet, and DepressionEmotion provide helpful data, but they are restricted in certain contexts, especially when applied across cultures [33], [40]. Most importantly, the absence of sociological set of labels restrained to mental health applications is a major limitation [33].

Multimodal Integration Complexity: Merging various modalities such as facial images and videos, speech samples, written text, and biological signals for emotional recognition is particularly difficult because of variation in data types, timestamping and methods for extracting features. Real-time clocking of different modalities without losing the advantages of each of them is very resource demanding and frequently involves compromise between accuracy and resource utilization [24].

Complexity in Emotion Representation: Emotions can be hard to distinguish because there is a high correlation and automation amongst them. Therefore, it can be difficult to set boundaries with individual emotions. If we recall the dimensional models or the valence-arousal together with the models that are categorical such as happiness, anger, fear, sadness, etc, they are quite often not suitable for use u and systems with mixed emotions or when looking for emotions that are ambiguous. For example, “calm” versus “bored” is a constant debate in several of the

current models whereas the distinguishment of “angry” and “frustrated” has now even become a trademark [27], [41].

Real-World Application Constraints: In comparison to controlled environments, ERS tend to not work as well when taken into real-world contexts. Things such as lighting, background noises, facial obstructions and speaker accent all impact the overall accuracy [22], [23]. QCNNs and other hybrid advanced systems have been helpful in trying to rectify some of these issues but are unfortunately complex to implement and work with [26].

Bias, Privacy and Ethical Concerns: When using emotional recognition systems for tasks, especially in healthcare and education, there are still significant ethical, privacy, and bias issues. Such systems use datasets that lack diversity which culminate in biased models that are unfair to a large population and raise issues of accountability and fairness in society. For instance, using biased algorithms can lead to misclassification in the interpretation of emotions for people who speak a particular language or are from a given culture [23], [40]. Privacy infringement can also be a key barrier to the widespread use of ERS since these systems extract sensitive data like speeches, text and even facial expressions. The danger of abuse, data leaks, and profiling constitute significant risks to user privacy and confidence, especially in the case of mental health applications. Additionally, the lack of accessibility to countless AI models leaves the user in the dark regarding the data processing procedures which in turn foster a lack of transparency [32], [41].

Tackling these limitations is essential to ensure the system’s reliability and applicability in different social contexts by predicting precisely the correct band of emotions.

2.1.4 Emotional Recognition Systems in Mental Health

Emotional Recognition Systems are gaining traction because of their exceptionally advanced methods of analyzing mental conditions. They use data from facial expressions, speech and text data to improve and help in mental health diagnosis, monitoring and providing further care support for emotional imbalance [20], [23].

Through the utilization of facial emotion recognition, professionals are now able to determine subtle emotional cues, which can assist in identifying complicated metal health conditions like depression and anxiety. Interpreting expressions and recognizing changes in emotion can be crucial, since detecting mental illness early and provide the correct personalized treatments can be the difference between life and undesired outcomes [20].

Another important component of emotion recognition systems is the analysis of speech and its vocal patterns – tone and intonation – to evaluate emotional well-being. This technology is very helpful for non-invasive diagnostics as it enables the identification of emotional distress through normal conversations [29], [42].

On the other hand, text sentiment analysis works great in texts such as therapy transcripts, Twitter or Facebook posts, etc. It helps in deriving context based on emotions and in noticing

thought and mood patterns that may indicate mental health disorders which may not be evident otherwise. This helps in improving the overall management of individuals suffering from mental disorders [33], [43].

The combination of these methods gives an accurate assessment of the person's internal and external emotional state. Such integration improves the precision of measurements and assists in the design of individual specific mental health treatment plans. This indicates the increased potential of technology to promote mental health recovery [20], [25].

2.2 Systematic Review

This systematic review aims to present a broad analysis of existent and relevant literature to this study and to answer a research question. The review attends defined systematic reviews patterns, comprising search terms, inclusion and exclusion criteria, and assessment of the selected studies quality, safeguarding reliable basis for the obtained results.

2.2.1 Methodology

A systematic review determines how to answer pre-defined research questions using explicit, reproducible methods to identify, critically appraise and combine results of primary research studies [44]. This review will follow the PRISMA methodology to increase the clarity, transparency, quality and value of this sub-section [10], providing insights on how to develop an emotion recognition system.

Based on its guidelines, this study will use the following metrics: Formulate appropriate research questions to answer the central question of this study. Retrieve relevant studies, using a search query, detailing the employed terms and the data sources used. Screen and decide on the defined exclusion and inclusion criteria. Assess the quality of the selected studies using established quality assessment standards. Extract the analyzed relevant data. Summarize the evidence, interpret the results, and present the findings to address the proposed research questions.

2.2.2 Research Questions

The primary goal of this research is to develop an emotion recognition system that helps professionals on mental health diagnoses by giving personalized therapeutic interventions. Therefore, focus on metrics such as facial analysis, speech and voice recognition, text sentiment analysis and physiological signals.

RQ1: What are the most effective methods for implementing emotion recognition technologies in mental health diagnostics?

RQ2: How do multimodal emotion recognition systems compare to single modal systems in terms of metrics for mental health applications?

RQ3: How effective are real-time emotion recognition systems in supporting personalized therapeutic interventions?

2.2.3 Data Sources

This review's selected and used studies, were retrieved from four electronic databases: PubMed[1], Academic Search Complete[2] and Web of Science[3]. PubMed is maintained by the Nacional Center for Biotechnology Information, renowned for its biomedical literature, life sciences and health-driven studies. Academic Search Complete, on the other hand, is provided by EBSCOhost, and is an inclusive scholarly database offering a large range of academic resources across multiple disciplines, including social sciences, humanities, and STEM fields (Science, Technology, Engineering, and Mathematics). Finally, Web of Science, a Clarivate Analytics platform, is a multidisciplinary research database known for its broad coverage of scientific, technical, and social sciences literature.

2.2.4 Search terms

The selection of the search query for the databases focused on identifying terms or keywords relevant to the research domain. The goal was to deepen the understanding of emotion recognition systems, along with the metrics and methodologies employed:

Mental Health OR Psychiatric Disorders OR Mental Disorder: the goal was to find studies that explore mental health diagnoses.

AND

Emotion Recognition OR Emotion Analysis OR Emotion Detection: these expressions were meant to obtain papers related to the use of emotion detection.

AND

Artificial Intelligence OR Deep Learning OR Machine Learning OR Computer Vision: the objective was to fetch AI-based researches.

[1] https://pubmed.ncbi.nlm.nih.gov/
[2] https://www.ebsco.com/products/research-databases/academic-search-complete
[3] https://www.webofscience.com/

2.2.5 Inclusion and Exclusion Criteria

The inclusion criteria utilized to filter the search query result was:

Table 1 – Inclusion Criteria

Inclusion Criteria	
IC1	Focuses on the application of AI to recognize and analyze human emotions for mental health support.
IC2	Explores the integration of emotion recognition technologies with real-world applications.
IC3	Assesses the effectiveness of real-time emotion recognition tools in improving mental health diagnostics and personalized care.
IC4	Provides studies or data evaluating the impact of AI-driven emotion recognition systems on mental illness diagnoses.

On the other hand, the following exclusion criteria were used to exclude sources from the selection process:

Table 2 – Exclusion Criteria

Exclusion Criteria	
EC1	Sources not written in English
EC2	Sources published before 2020
EC3	Sources besides journal articles, chapters, conference, proceedings, and books
EC4	Sources that do not present studies related to the use of emotion detection on mental health diagnoses
EC5	Sources that present studies that leverage wearable devices to detect emotions
EC6	Duplicated sources

2.2.6 Quality Assessment

After the papers were assessed by the inclusion and exclusion criteria, its quality was verified regarding:

- The relevance of the study for the mental health field.
- The clarity of the results. Each paper was evaluated whether it presents clear metrics,
- solid discoveries, and definitive conclusions.
- The number of citations the paper has.

2.2.7 Data Extraction and Synthesis

The search for articles was concluded with a sum of 260 sources, within all utilized databases. After the removal of the duplicated sources, 184 papers were the target of a title and abstract screening process. In that phase, the sources were assessed whether it met the inclusion or exclusion criteria. That led to a discard of 115 articles, to a re-evaluation of 25 potentially relevant articles and 44 as relevant. The potentially relevant sources went through an extra results and conclusions evaluation procedure, resulting in a total of 8. In this stage the 52 relevant papers considered went through a quality assessment full-text screening process, to analyze its relevance to the current research question. This selection process obtained a total final of 27 sources (cf. Figure 2), which were selected for the results of this review.

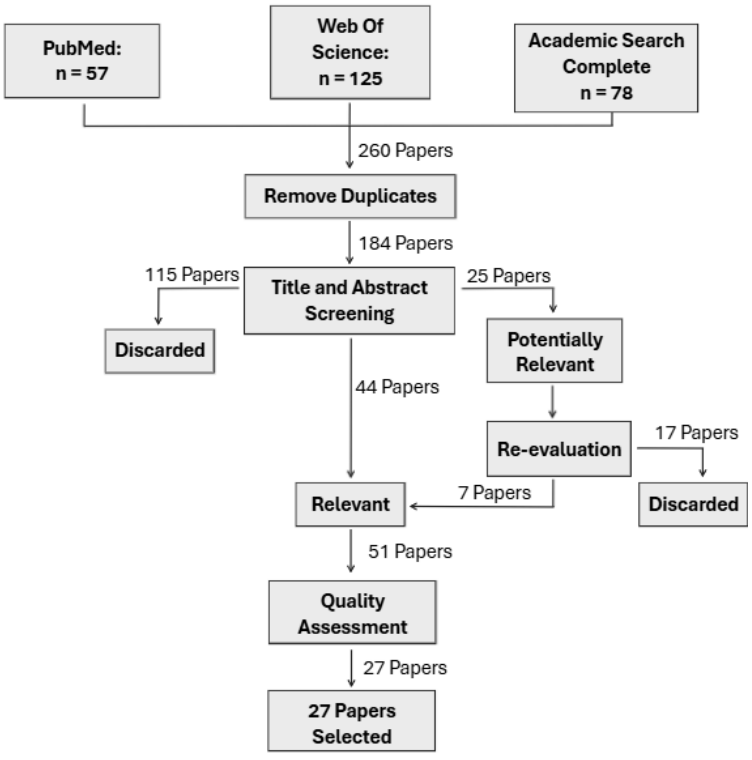

Figure 2 – PRISMA Flow Diagram

2.2.8 Research Questions Answers

RQ1: What is the most effective methodology for implementing emotion recognition technologies in mental health diagnostics?

Three main methodologies to implement the emotion recognition system were found in the papers. Following, each technique will be presented separately.

Facial Expression Emotion Detection

Facial expression emotion detection for mental health has been extensively studied through various deep learning and artificial intelligence methodologies, aiming to improve the accuracy and efficiency of mental health diagnostics. Across multiple studies, common themes emerge, particularly in leveraging deep neural networks, multimodal approaches, and novel computational techniques to enhance emotion recognition.

Deep learning-based approaches remain dominant in emotion recognition, with CNN architectures frequently employed for feature extraction. Fei et al. [22] utilize AlexNet's Fully Connected Layer 6 to extract deep features, followed by a Linear Discriminant Analysis classifier for emotion classification. Their model, tested on JAFFE, KDEF, CK+, FER2013, and AffectNet, achieves superior classification accuracy, particularly on CK+ and AffectNet, outperforming traditional handcrafted feature extraction methods. Similarly, Peng et al. [15] validated on a self-built dataset collected from college students, an integration of LSTM networks into CNN architectures to improve temporal learning, allowing for better recognition of emotions across time sequences which proves to be very promising in real-time systems.

Hossain et al. [26] introduce a Quantum Convolutional Neural Network to enhance feature extraction efficiency by leveraging quantum computing principles. Their method significantly reduces training time while maintaining high accuracy when tested on KDEF, Static Facial Expressions "in the Wild" (SFEW 2.0), and FER2013 datasets. The integration of quantum computing in deep learning models shows promise in making real-time facial emotion detection more efficient. Meanwhile, Jain et al. [40] address the challenge of limited labelled data by proposing EM-UDA, which improves emotion classification in cross-dataset scenarios. By training on AffectNet and CK+ while testing on FER2013, their model achieves an accuracy of 83.9%, proving the effectiveness of domain adaptation in handling dataset shifts and improving generalization.

Several studies emphasize the importance of integrating multiple modalities to improve emotion detection accuracy. Wu et al. [25] incorporate Laban Movement Analysis (LMA) into their study on bodily expressed emotions, using the Body Motor Elements (BoME) dataset to analyze movement-based emotional expressions. Their findings confirm that movement-based emotion recognition enhances classification accuracy when integrated with facial recognition data. Similarly, Li et al. [18] combine facial emotion recognition with psychological scales, such as DASS-21, to improve diagnostic efficiency in large-scale mental health screenings. Their

model, pre-trained on FER2013 and CFEE, shows that combining facial recognition with psychological assessment tools leads to more reliable mental health evaluations.

Xu et al. [14] and Shanthi et all. [21] further explored multimodal recognition by integrating speech-based emotion analysis, using RAVDESS, TESS, SAVEE, and CREMA-D datasets. Findings from these studies indicate that combining speech and facial expression data improves classification accuracy compared to unimodal models, though the exact percentage increase is not explicitly stated. This aligns with studies that argue for the necessity of incorporating voice tone, body language, and contextual text analysis to improve the robustness of emotion detection systems.

Further advancements have been made by integrating cross-attention and feature recalibration mechanisms into deep learning frameworks. Zhou et al. [20] propose an interaction–rectification pair, consisting of a cross-attention block and an element recalibration block, which enhances emotion representation by embedding global contextual information into facial feature maps. Their approach effectively reduces emotion confusion and improves recognition precision, particularly in dynamic environments where facial expressions alone may not provide sufficient discriminative cues.

Despite the progress in facial expression emotion detection, several challenges persist, particularly regarding dataset limitations, model generalization, and computational efficiency. Many datasets, such as CK+ and JAFFE, contain posed facial expressions that do not generalize well to real-world applications, struggling to correctly identify the expressed emotions on poor lighting conditions, bad image quality and background complexity, which reduces the performance in spontaneous emotion detection [22]. Xu et al. [14] acknowledge that AI-driven mental health assessments face challenges related to demographic biases in existing datasets, which can affect fairness and reliability across diverse populations. They also highlight ethical concerns such as privacy risks and the need for robust safeguards when deploying AI-based emotional recognition in mental health applications.

Speech and Voice Emotion Analysis

Adeleye et al. [45] propose a wavelet transform-based feature extraction technique for speech emotion recognition, focusing on mental health monitoring. Their study derives features from energy content and entropy measures to classify emotions such as fear, sadness, anger, anxiety, and disgust. Using convolutional neural networks, they demonstrate that their discrete wavelet transform-based method achieves an unweighted accuracy of 83.7% and a weighted accuracy of 81.7%, outperforming traditional FFT-based features such as Mel-Frequency Cepstral Coefficients (MFCC) and Mel-Spectrograms.

Singh et al. [42] propose a self-attention-based deep learning model that combines CNN and LSTM networks for speech emotion recognition. The model is designed to enhance emotion classification by leveraging spectral and rhythmic information, with experiments demonstrating that Mel Frequency Cepstral Coefficients are the most effective feature. Their method, tested on the RAVDESS, SAVEE, and TESS datasets, achieves an average test accuracy of 90%,

outperforming conventional deep learning models. The study highlights the importance of attention mechanisms in improving speech-based emotion recognition for potential applications in mental health monitoring.

Zhao et al. [28] introduce a multi-head time-dimension attention-based LSTM model for depression detection from speech. Their approach preserves the temporal relationships in speech sequences using selected frame-level features. By incorporating multi-head attention, they effectively project the LSTM output into different subspaces, enhancing classification accuracy. Their model achieves improvements of 2.3% and 10.3% over traditional LSTMs on the DAIC-WOZ and MODMA datasets, respectively. Their findings suggest that emphasizing emotionally salient regions in speech improves the detection of depression symptoms.

Xu et al. [14] proposed a dynamic multimodal feature recognition framework that integrates traditional psychological scales with machine learning-based emotion recognition for mental health assessment. Their study applies deep learning models to analyze both speech and facial expressions, classifying emotions related to stress, anxiety, and depression. By combining multimodal emotion recognition with conventional assessment scales, they demonstrate improved accuracy and reliability in screening mental health conditions among college students. Shanthi et al. [21] propose an integrated mental health multimodal assessment system combining Facial Emotion Recognition, Speech Emotion Recognition, and psychological scales. Their approach extracts audio features such as Mel-Frequency Cepstral Coefficients using the Librosa library and employs deep learning techniques for emotion classification. The study demonstrates that integrating multiple modalities — speech, facial expression, and self-reported psychological assessments — improves the accuracy of depression detection, with the FER model achieving 91% accuracy and the SER model 82%.

Caulley et al. [46] investigated AI-driven speech emotion recognition for pediatric psychiatric assessment, specifically in children with high Adverse Childhood Experience scores. Their study evaluates transformer-based and CNN-based models for classifying emotional states in structured speech samples. They find that transformer-based models achieve better precision and recall (86% and 79%, respectively) for binary emotion intensity classification, while CNNs generalize more effectively across multiple emotional states. Their findings suggest that AI-driven speech analysis can support clinicians in triaging youth for appropriate mental health interventions. Similarly, Alemu et al. [30] research also addresses the recognition of mental issues on children and adolescents. They developed a Gaussian Mixture Model based speech emotion recognition system that achieved 85.35% of accuracy when using emotion frame selection, emphasizing its potential for real-time classification.

Across these studies, the common challenges identified in emotion recognition for mental health applications include dataset limitations, difficulty in differentiating overlapping emotional states, and the need for noise-robust feature extraction. While deep learning-based approaches such as CNNs and LSTMs offer promising accuracy, they require large, high-quality labeled datasets to generalize well.

Text Emotion Recognition

Jiang et al. [24]emphasize the advantages of multimodal emotion recognition, integrating text, facial expressions, and physiological signals to improve accuracy. Their study highlights that single-modal approaches, including text-based analysis alone, face challenges in accurately capturing human emotions due to external noise, contextual variations, and the subjective nature of textual interpretation. They point out that multimodal fusion, such as combining speech, and facial expressions, can mitigate these challenges by providing complementary emotional cues.

Meshram and Rambola [16] further support this by proposing a deep learning-based framework for depression diagnosis. Their approach integrates facial feature extraction using convolutional neural networks with linguistic analysis employing k-nearest neighbors and Random Forest regression. While the study reports a 2.7% improvement in accuracy, this enhancement is primarily attributed to facial detection and feature extraction rather than text-based analysis alone. However, they highlight that user-generated textual descriptions play a crucial role in distinguishing between different depressive groups, supporting the need for advanced text emotion recognition models in mental health applications.

In contrast to Jiang et al.'s [24] multimodal approach, which integrates text, facial expressions, and voice analysis for emotion recognition, Oh et al. [43] focus on a purely text-based emotion recognition framework for diagnosing depression. Their machine learning model, utilizing text transcripts from routine psychiatric interviews, effectively classified depressive and non-depressive patients with an AUC of 0.85. The study found that emotion distribution in language is a strong indicator of mental health status, with disgust emerging as the most significant emotional marker differentiating depressive from non-depressive individuals ($p < 0.001$).

Similarly, Rahman et al. [33] underscore the potential of text-based emotion detection by introducing DepressionEmo, a dataset comprising 6,037 Reddit posts labeled for eight depression-related emotions. Their research applies deep learning techniques, including BERT and GAN-BERT, to classify emotions such as anger, hopelessness, and suicidal intent, achieving an F1 Macro score of 0.76. Among the classified emotions, suicidal intent showed the highest F1 Macro score, highlighting the dataset's relevance in detecting depression-related linguistic patterns.

While both Oh et al. [43] and Rahman et al. [33] focus on text-based models for mental health detection, Kodati and Ramakrishnudu [34] take this a step further by incorporating multi-task learning. Their approach utilizes soft-parameter sharing transformers to simultaneously detect mental health disorders, classify suicide-related texts, and identify emotions in suicide notes. Their research introduces three models — SPS-LSTM-AM, SPS-BiGRU-SAM, and SPS-BNN-MHAM— which integrate disorder detection, suicide classification, and emotion detection to enhance performance. The soft-parameter sharing technique allows for shared learning across related tasks while maintaining independent task-specific layers, leading to higher accuracy and generalization. Their models achieve exceptional performance across multiple mental health datasets, with reported accuracies of 96.9%, 97.4%, and 98% and F1 scores of 93.8%, 94%, and

94.6%, respectively. By integrating emotion detection in suicide notes as the primary task while using disorder and suicide classification as auxiliary tasks, their study demonstrates that learning multiple related tasks simultaneously can significantly enhance text-based emotion recognition performance.

Despite the advances in emotion recognition, integrating AI-driven sentiment and emotion analysis into mental health support systems raises ethical concerns and even hardware problems, as mentioned by Dheeraj et al.[32] that emphasizes model training can be computationally exhaustive. Likewise, Denecke and Gabarron [41] also discussed risks associated with AI training such as bias in sentiment analysis models, the potential misinterpretation of user emotions, and privacy concerns in handling sensitive mental health dat. Their research highlights that AI models trained on general sentiment datasets often fail to accurately represent individuals with depression, leading to misclassification and inappropriate interventions. This is particularly concerning chatbot-based interventions, where misinterpretation of emotional states may result in ineffective or even harmful responses. Their study underscores the importance of ethical considerations in AI-driven mental health applications, aligning with the concerns raised by Jiang et al. [24] regarding the limitations of automated emotion detection. Similarly, Meshram and Rambola [16] warn that deep learning models integrating facial and linguistic features require careful design to avoid biases that could negatively impact mental health diagnostics. Denecke and Gabarron [41] argue that ensuring transparency, obtaining informed user consent, and embedding AI models within supervised clinical settings are essential steps to mitigate ethical risks.

The implementation of emotion recognition technologies in mental health diagnostics requires a nuanced approach, as each methodology has its own strengths and limitations. Facial expression emotion detection, leveraging deep learning models like CNNs and LSTMs, has demonstrated high accuracy in controlled settings but struggles with real-world generalization due to dataset biases, lighting conditions, and spontaneous expressions. Speech and voice emotion analysis, particularly through wavelet transform-based feature extraction and attention-based LSTM models, provide valuable insights into emotional states through tone, rhythm, and frequency. However, it faces challenges in distinguishing overlapping emotions and handling noise variations. Text-based emotion recognition, powered by deep learning models such as BERT and multi-task learning frameworks, has shown promise in detecting mental health indicators from written content. Nonetheless, its effectiveness is often limited by the subjective nature of text interpretation and potential ethical concerns, including bias and misclassification.

Given these individual strengths and weaknesses, multimodal approaches stand out as the most effective methodology for implementing emotion recognition in mental health diagnostics. By integrating facial expression analysis, speech emotion recognition, and textual sentiment detection, multimodal systems can compensate for the vulnerabilities of single-modal approaches, leading to improved accuracy and robustness.

RQ2: **How do multimodal emotion recognition systems compare to single modal systems for mental health applications?**

Multimodal emotion recognition systems regularly outperform unimodal approaches in key evaluation metrics such as accuracy, precision, recall, and F1-score since, according to Jiang et al. [24], single-modal systems often struggle with the low emotional information available which affects directly the model metrics. Zou et all. [20] point it out that these systems leverage complementary modalities to reduce false positives and false negatives, cross-validating emotional indicators across different sources, thereby providing a holistic view of an individual's mental health. Their study proposes a multimodal system and seeks to avoid single-modal main problems such as the easiness that can be influenced by various noises and cannot fully reflect an accurate detection of emotional states. This multimodal approach allows for improved precision and recall, as contradictory signals from one modality can be adjusted by corroborating evidence from another. Lastly, their experiments demonstrated how cross-attention mechanisms improve accuracy, confirming that integrating multiple input streams leads to better generalization and resilience against environmental noise.

Xu et al. [14] study presents a novel approach to mental health assessment that combines interactive questionnaires with multimodal intelligent recognition technology. By integrating data from facial emotion recognition, audio emotion analysis, and self-reported psychological scales, their method enhances the precision and reliability of mental health condition predictions. Experimental results from a study of 1,500 college students demonstrate that the multimodal emotion recognition model achieves 90.2% accuracy in audio emotion recognition and 82.3% accuracy in video emotion recognition, stating that reliability is higher than traditional unimodal models, that only rely single modal data.

On the other hand, Adeleye et al. [45] study focuses on unimodal emotion recognition from speech using wavelet transform features. The authors achieved an unweighted accuracy of 83.7% and a weighted accuracy of 81.7% in detecting emotions like fear, sadness, and anger from speech signals. While these results are promising, the study notes that the single-modal approach struggles with environmental noise and limited emotional cues, making it less robust compared to multimodal methods that incorporate complementary data sources.

Similarly, Hossain et al. [26] study explores a unimodal facial expression recognition system for mental health analysis using a Deep Quantum Convolutional Neural Network. The model is designed to detect emotions from static and sequential facial images extracted from medical healthcare datasets. While achieving 81.95% accuracy during training, its performance may drop, since the authors highlight that relying solely on facial expressions can lead to misclassification due to variations in distortions, noise, and subtle emotional expressions, factors that multimodal systems can better handle by integrating complementary modalities such as speech and text sentiment analysis.

Regarding textual emotion detection, Kodati et al. [34] explores a unimodal text-based emotion analysis approach for detecting mental health conditions. The authors propose a multi-task learning model that uses soft parameter sharing transformers to analyse psychiatric texts and

social media posts. Their model achieves 96.9% and 97 97.4%, and 98%, with F1 scores, precision, and recall all above 93%, demonstrating their ability to capture emotional and psychological patterns in various texts. However, the study notes that text-only models lack contextual depth, as written expressions may not always reflect true emotional states, making them less reliable compared to multimodal approaches that integrate facial and vocal cues.

Zhu et al. [17] research, compares single-modal and multimodal approaches using voice and video modalities for emotion recognition. Although, their multimodal model outperformed the unimodal one by approximately 11%, the authors leave some warnings regarding multimodal systems. Like unimodal systems, multimodal models can also experience problems such as imbalance on the different used datasets, where certain emotions have fewer samples, leading to biased recognition and lower accuracy for underrepresented emotions. Additionally, feature extraction and fusion challenges can introduce redundancy making the model less efficient. The high computational cost of deep learning architectures like Attention-LSTM and VGG-16 may also hinder real-time performance. Lastly, the fusion of the different modalities can introduce inconsistencies majority observed when each model makes different classifications. This explains why sometimes unimodal systems outperform systems based on various methodologies in terms of accuracy.

The comparison between multimodal and unimodal emotion recognition systems for mental health applications reveals that while multimodal approaches generally offer advantages, they are not without challenges. Unimodal models, whether based on speech, facial expressions, or text sentiment analysis, often struggle with environmental distortions, limited emotional cues, and difficulties in interpreting nuanced expressions, leading to potential misclassifications. Multimodal systems, by integrating diverse data sources, provide a more comprehensive and reliable analysis. However, they also introduce issues such as synchronization difficulties, increased model complexity, and potential redundancy between modalities. Despite these challenges, multimodal frameworks tend to outperform unimodal ones in most cases, making them more effective for mental health assessments and real-world applications.

Table 3 – Studies Modalities and Respective Metrics

Author	Modality	Accuracy	Precision	Recall	F1 Score
Jiang et al. [24]	Visual (Facial), Audio, Text	77.6%	-	-	-
Zhou et al. [20]	Visual (Facial and Context)	83.76%	-	-	-
Xu et al. [14]	Visual, Audio	86,23%	-	-	-
Adeleye et al. [45]	Audio	72,32%	-	-	--
Hossain et al. [26]	Video	78.70%	56.34%	61.55%	61.55%
Kodati et al. [34]	Text	96%	> 94%	> 93%	93.1%
Zhu et al. [17]	Visual, Audio	78.5%	-	-	-

RQ3: **How effective are real-time emotion recognition systems in supporting personalized therapeutic interventions?**

Early detection of mental illnesses is crucial for effective treatment and intervention, and this is where real-time emotion recognition systems play a vital role. Leung et al. [47] investigated how real-time emotion recognition systems can prove to be promising applications in clinical settings by providing immediate and adaptive responses to patients' emotions. Their study evaluated an AI-based co-facilitator for online cancer support groups, which monitored participant distress in real-time through text-based emotion analysis, helping therapists adjust their approach dynamically. The study found that such AI-assisted tools improve patient engagement and enable better emotional support.

A major area of application is in college mental health screening and diagnostics. Li et al. [18] proposed an intelligent mental health identification method that combines facial emotion recognition with psychological questionnaires. Their system analyzed facial expressions in real time to detect early signs of stress, anxiety, and depression among students. The study demonstrated that real-time AI-assisted screening improves efficiency in large-scale psychological assessments, reducing the workload for mental health professionals.

Multimodal emotion recognition systems have also been studied as a means to increase accuracy in real-time ERS. Jiang et al. [24] multimodal fusion techniques that combine facial analysis, vocal emotion recognition, and text-based sentiment detection. Their findings suggest that integrating multiple data sources leads to more robust emotion detection models, minimizing false positives and negatives. These systems, when applied to mental health diagnostics, enable early intervention by detecting subtle emotional cues that single-modal approaches may overlook.

Additionally, the study on machine learning algorithms and feature sets for automatic vocal emotion recognition in speech. Dõgdu et all. [29] examined different deep learning models for processing real-time speech-based emotions. The research found that Support Vector Machines, and Convolutional Neural Networks were effective in classifying emotional states from live speech data. Their findings support the application of speech-based ERS for personalized therapeutic interventions, where continuous monitoring of vocal tone and pitch variations can detect early signs of emotional distress.

Another key consideration is the real-time ethical and privacy implications of real-time emotion recognition in psychiatric applications. Aguilera et al. [31] proposed a blockchain-based AI model designed to secure patient data in real-time ERS. Their research highlights that real-time data processing in mental health applications requires constant monitoring of sensitive emotional states, which raises concerns regarding data security and patient confidentiality. By integrating blockchain and federated learning, their model enables decentralized real-time emotion tracking, allowing for secure and tamper-proof storage of emotional data, while ensuring fast and reliable processing without centralized vulnerabilities.

Similarly, Aina et al. [23] explored the necessity of explainable AI models for real-time psychiatric decision-making. Their study found that real-time emotion recognition models, when used in therapeutic settings, must not only detect emotions accurately but also provide transparent and interpretable reasoning for clinicians. The ability to understand and verify the AI's decision-making process in real time is critical for therapists to make informed clinical judgments and prevent potential misinterpretations of emotional states. Their findings suggest that enhancing interpretability in real-time emotion recognition AI models ensures that the outputs can be utilized and integrated into psychiatric assessments.

These studies emphasize that for real-time ERS to be widely adopted in mental health settings, they must incorporate robust privacy safeguards and real-time explainability mechanisms. Without these protections, the continuous monitoring of emotional states in real-time could pose risks of data breaches, ethical concerns, and clinician distrust in AI-driven decisions.

2.3 Discussion

The analysis of the state of the art highlighted several key insights that directly support the development of an emotion recognition system for mental health applications. The reviewed studies covered different methodologies, evaluation metrics, common challenges, and real-world applications, enabling a critical comparison of their contributions and limitations.

Regarding the methodologies, three main approaches—facial expression analysis, speech and voice emotion recognition, and textual sentiment analysis—were found to be widely used. Facial analysis techniques based on deep learning models (such as CNNs, ResNet, and QCNNs) have achieved high accuracy under controlled conditions, but face difficulties in generalizing spontaneous emotions and real-world environments due to dataset bias, lighting, and noise. Speech-based methods can capture emotional tone and physiological changes effectively and are non-intrusive, but struggle with background noise and overlapping emotions. Text-based emotion recognition can detect linguistic markers of mental health disorders yet is affected by subjectivity and context ambiguity. These findings indicate that single-modal approaches alone are insufficient to robustly detect complex and nuanced emotional states.

Concerning evaluation metrics, accuracy, precision, recall, and F1-score are the main indicators of model performance, while ROC-AUC and standard deviation offer complementary perspectives on robustness and generalization. The reviewed literature shows that unimodal models may achieve high performance on specific datasets but often suffer from lower recall and reduced generalization when applied to heterogeneous real-world data. This reinforces the need for models that not only perform well in controlled settings but also maintain reliability across diverse scenarios.

On the topic of challenges, the main barriers identified include limited and culturally biased datasets, complexity in fusing different modalities, ethical and privacy risks, and the gap between laboratory settings and real-world conditions. These challenges highlight that

technical performance alone is not enough; ethical, transparent, and explainable models are essential for safe integration into clinical contexts.

With respect to applications in mental health, the literature demonstrates that emotion recognition can support the early detection of conditions like depression and anxiety by capturing subtle emotional cues in facial expressions, speech patterns, and written language. Importantly, studies show that combining these modalities improves diagnostic accuracy and reliability, offering a more complete picture of the patient's emotional state. This aligns with the objective of this work, which seeks to create a system capable of supporting personalized therapeutic interventions.

The systematic review results further support this conclusion: while unimodal models present strong isolated performances, multimodal systems consistently outperform them in accuracy, recall, and robustness, as they cross-validate emotional indicators from different sources and compensate for missing or noisy data. Additionally, real-time emotion recognition emerged as a crucial factor for clinical usability, as it enables adaptive responses during therapy and mental health monitoring, enhancing patient engagement and treatment outcomes.

In summary, the discussion of the state-of-the-art leads to three main conclusions:

1. Multimodal emotion recognition systems are more effective than unimodal ones for mental health diagnostics, as they provide richer and more reliable emotional context.
2. Real-time processing capabilities are essential to support adaptive and personalized therapeutic interventions.
3. Ethical, privacy and explainability mechanisms must be integrated from the design stage to ensure clinical applicability and user trust.

Based on these conclusions, this thesis will pursue the development of a multimodal, real-time, explainable emotion recognition system that combines facial analysis, speech and voice emotion recognition, and textual sentiment analysis to support mental health professionals in diagnostic and therapeutic decision-making.

3 Methods and Materials

This chapter presents the methods, tools and materials employed during the development of the project, followed by an explanation of the procedures utilized to perform the system evaluation and experimentation. Additionally, it displays the selected datasets, and the preprocessing steps applied to it. Finally, it discusses ethical, security and data protection factors that were considered.

3.1 Method and tools

After state-of-the-art research and a systematic review to investigate how to best apply emotion detection to mental health diagnoses. The integration of facial expression analysis, speech and voice emotion recognition, and textual sentiment analysis aims to enhance the model accuracy and precision in detecting emotional states. The following sections describe the methods, tools, and datasets used, along with its preprocessing, system evaluation procedures, and ethical considerations.

Several datasets were mentioned in the state-of-the-art chapter and, within them, the chosen datasets for this project have been strictly selected to have same emotions categories to ensure models detections accuracy. The FER2013 dataset is broadly used in state-of-the-art facial emotion recognition research, mostly due to its inclusive emotion categories, standardized image format, and strong community support. All its images are 48×48-pixel grayscale, ensuring uniformity and simplifying preprocessing. Additionally, the dataset includes seven different emotion labels – Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral– providing a broad range for model training [21], all motives why this dataset was chosen.

Regarding speech and voice emotion recognition, the datasets CREMA-D, RAVDESS, SAVEE, and TESS, are strongly utilized for their thorough coverage of emotional expressions, speaker cultural diversity, and high-quality recordings. Therefore, the utilization of all these datasets

can improve the model's ability to generalize through diverse emotional traces and speaker variations, which contribute to a more robust and accurate emotion prediction. This approach aligns with state-of-the-art practices in the field where leveraging diverse datasets is proven to be beneficial to developing models capable of understanding and interpreting the complex nature of human emotions in speech.

Furthermore, after conducting research, it was identified a suitable dataset for textual emotion detection component of this thesis. The dataset utilized consists of anonymized conversations between patients and experienced psychologists. Similar datasets were mentioned in the state-of-the-art research, including suicide notes and social media posts.

Lastly, the datasets will go through a preprocessing phase to ensure data quality and relevance. This includes cleaning the data, normalizing text data and feature extraction. After that, the proposed model will be trained, tested and evaluated.

3.2 Datasets

This section presents the characteristics of the chosen datasets for each emotion detection methodology – Facial, Speech and Textual emotion analysis – by employing an exploratory analysis. This analysis examines, for each dataset, its data structure, field types and key attributes. Additionally, the process of preprocessing each dataset is described, utilizing techniques of data transformation and preparation; a pipeline for each emotion analysis methodology is exhibited, detailing the flow, processes and transformations that the datasets experienced before being fed into the model.

Furthermore, Figure 3 illustrates the entire system workflow, starting with request for consent from the patient. The session only starts when the patient agrees to the use of the developed application. When the consultation begins, the changes in the patient's emotions are assessed through facial expression analysis, allowing the therapist to monitor emotional developments in real time and tailor a more personalized intervention. Once the session ends, the consultation recording is processed, and an analysis of the patient's emotional changes on three levels (facial expressions, voice patterns, and interactions/responses) is carried out. The data processed by the three emotion models is then analyzed by an LLM, which is responsible for generating a detailed consultation report, providing information about sudden changes in emotion and their triggers, as well as recommendations for therapeutic follow-up and future interventions. Finally, the therapist can "interact" with the report and the consultation through a chatbot that has access to all the meeting's refined data, enabling immediate clarifications and deeper insights into the patient’s emotional state shifts and behaviors.

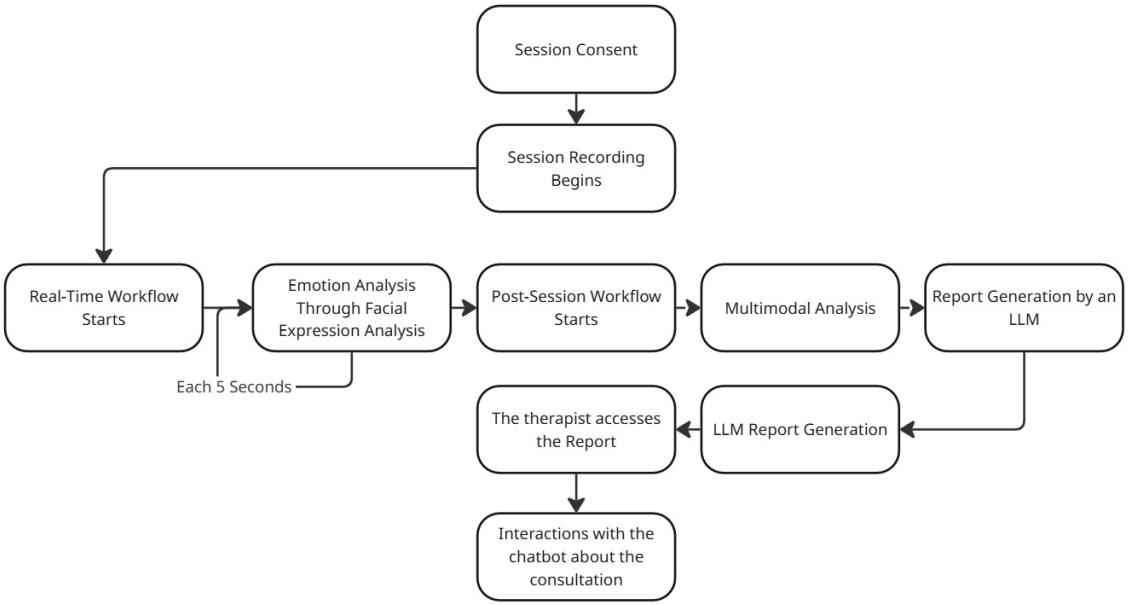

Figure 3 – System's Pipeline

3.2.1 Facial Expression Emotion Detection

The research conducted to identify a suitable dataset to detect emotions through facial expressions led to the discovery of FER2013. The images in this dataset were already distributed on a CSV format, as represented on Table 4. Each row represents an image illustrated by 3 columns – "emotion", "pixels" and "usage". The first column is relative to the emotion displayed by the picture, which are represented by integer numbers between 0 and 6, corresponding to the seven emotions that were previously referred. The "pixels" column contains the image pixels values in a grey scale and with a 48x28 pixels size. Finally, the "usage" indicates if the correspondent image belongs to the training or test set.

Table 4 – FER2013 dataset data example

emotion	pixels	usage
0	70 80 82 72 58 58 60 63 54 58 60 48 89 115 121 119 115 110 98 91 84 84 90 99 110 126 143 153 158 171...	Training
2	231 212 156 164 174 138 161 173 182 200 106 38 39 74 138 161 164 179 190 201 210 216 220 224 222 218...	Training
6	60 66 78 66 65 90 90 51 23 14 16 13 19 27 42 56 66 77 83 86 92 98 112 129 135 131 132 127 123 113 107 97 …	PublicTest

The following preprocessing process (cf. Figure 4) was applied to generate the final dataset used to train and test the proposed model:

1. **Shuffling the Data:** The shuffling ensures that the data is randomly distributed, preventing bias that may appear form the data ordering.
2. **One Hot Encoding:** Since the emotion labels are categorical data, their structure needs to be converted into one-hot encoded vectors to be a suitable input for the model.
3. **Standardization:** To normalize the pixels data a standardization is applied to have a mean of 0 and a standard deviation of 1, which helps the model converge faster during training.
4. **Reshaping the Data to (48, 48):** The pixel data is reshaped back into 48x48-pixels format.
5. **Train-Test-Validation Split:** The dataset is split into three parts training, test and validation.
6. **Data Augmentation:** To increase the variety of training data to prevent overfitting, data augmentation is employed. This includes shifting, flipping, and zooming the images during training.

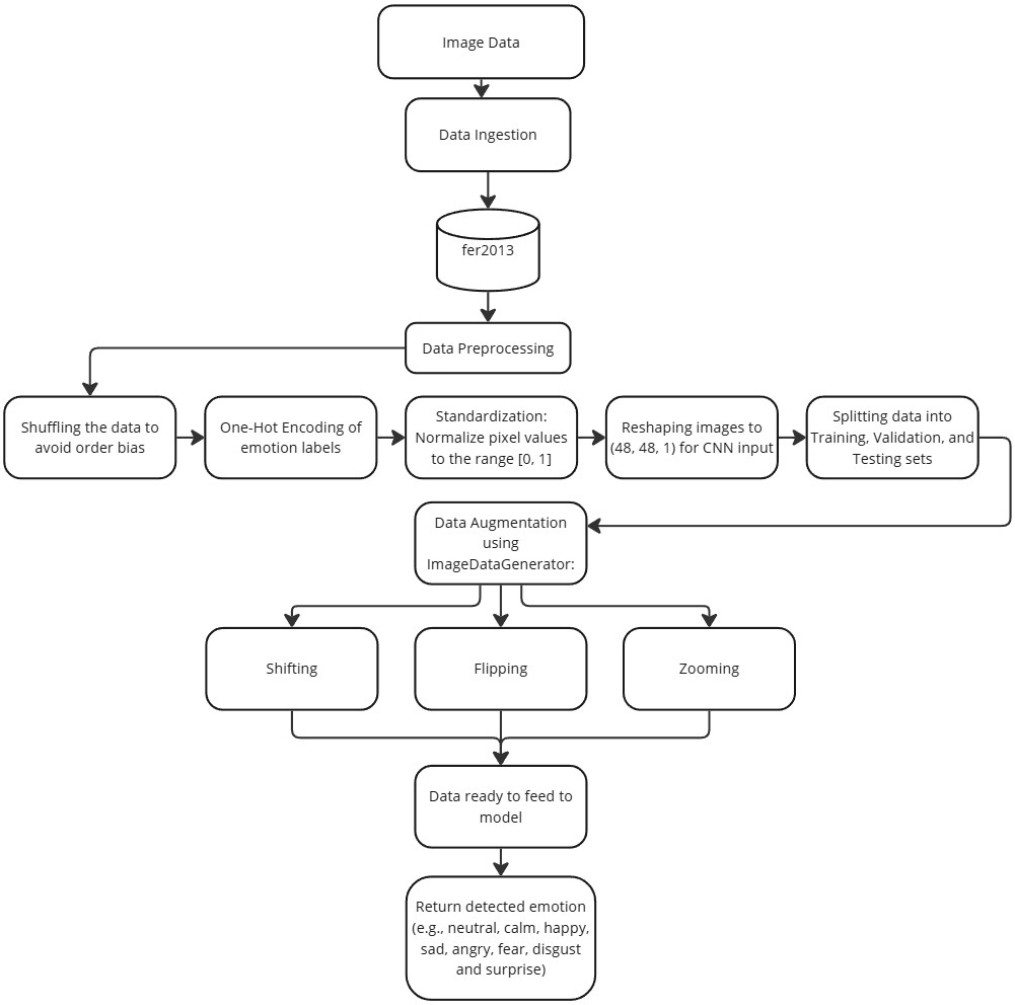

Figure 4 – Facial Expression Preprocessing Dataset Flow

3.2.2 Speech and Voice Emotion Analysis

The data supplied for the speech and voice emotion analysis model input had origin in 4 distinct datasets – Crema-D, Ravdess, Tess and Savee. All four datasets contain different audio files that correspond to a determined emotion. These audios were further converted in data frames with the following structure: “emotions” and “path”. The “emotions” column contains the emotion of each audio represented on the “path“ column (cf. Table 5).

Table 5 – Speech datasets data example

emotions	**path**
disgust	1001_DFA_DIS_XX.wav
angry	1001_DFA_ANG_XX.wav
happy	1001_DFA_HAP_XX.wav

The preprocessing employed involves the following steps (cf. Figure 5):

1. **Data Augmentation**: Apply small perturbations to the original voice signal helps improve model’s generalization.
    a. Noise Injection - Add random Gaussian noise to the audio signal. This allows the model to learn to distinguish the actual signal from noise.
    b. Stretching (Time Stretching) - Changes the speed of the audio without affecting its pitch. This helps the model generalize to variations in speech rate since audio speed signals can vary from person to person.
    c. Shifting - Randomly shifts the audio signal forward or backward in time. Safeguards the model against variant temporal position of features.
    d. Pitch Shifting - Changes the pitch of the audio without affecting its speed. It’s applied since different speakers have different vocal characteristics.
2. **Feature Extraction:** Since audio signals cannot be directly fed into models, feature extraction obtains meaningful features that the models can recognize.
    a. Zero Crossing Rate (ZCR) – To measure the r ate at which the audio signal changes sign. Important to identify the noisiness and spectral properties of the signal.
    b. Chroma Short-Time Fourier Transform (Chroma STFT) - Represents the 12 distinct pitch classes in music based on the harmonic content of the audio. Relevant to capture melodic and harmonic information.
    c. Mel Frequency Cepstral Coefficients (MFCC) - Represents the short-term power spectrum of the audio. Good for Capturing timbral and textural information.
    d. Root Mean Square (RMS) Value - Measures the average energy or loudness of the audio signal. Provides information about the intensity of the signal.
    e. Mel Spectrogram - Represents the power spectrum of the audio in the mel scale, which approximates human auditory perception. Captures both frequency and temporal information.

3. **Data preparation:**
    a. Label Encoding - The target labels (emotions) are one-hot encoded using One Hot Encoder to handle the multiclass classification problem.
    b. Train-Test Split - The dataset is split into training and testing sets.
    c. Feature Scaling - Features are normalized using Standard Scaler to ensure all features have zero mean and unit variance.
    d. Reshaping for Model Compatibility - The feature vectors are reshaped to add an additional dimension (to make it compatible with the model).

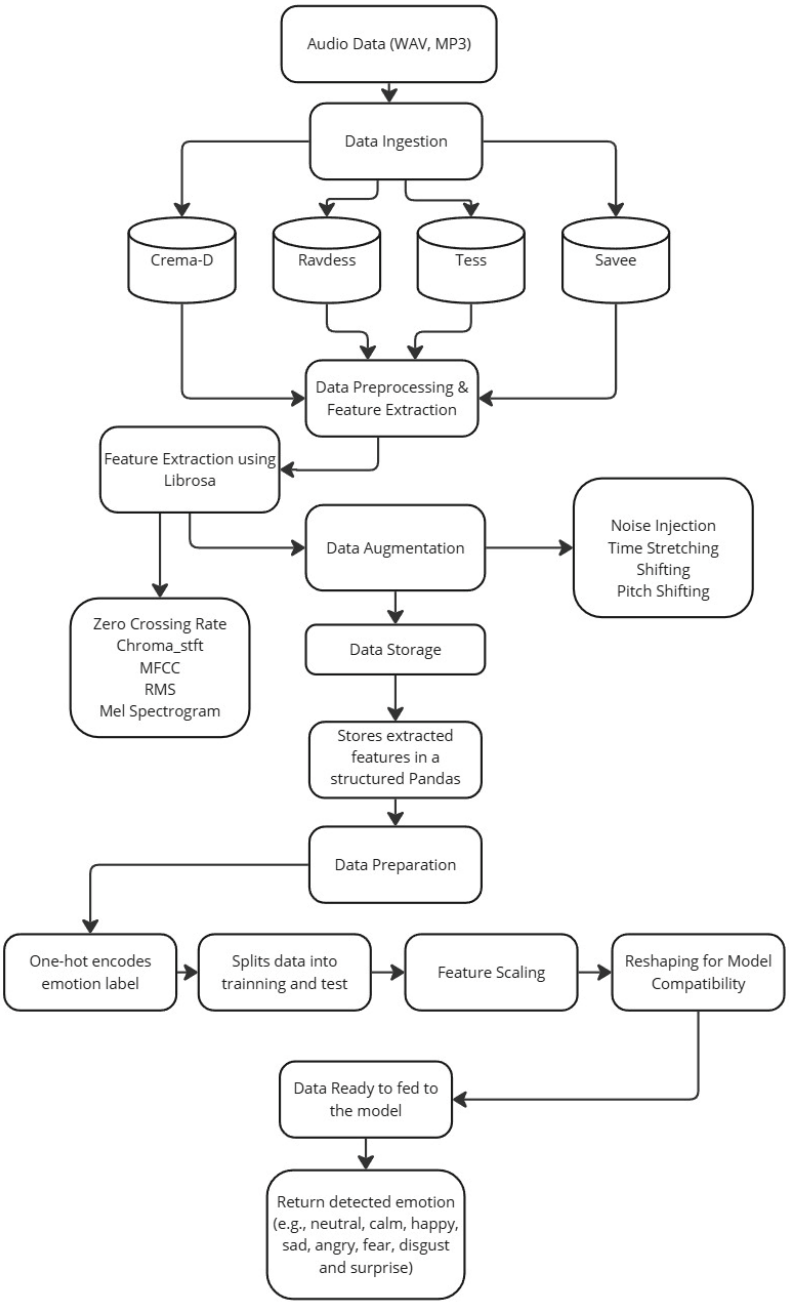

Figure 5 – Speech Emotion Preprocessing Dataset Flow

3.2.3 Text Emotion Recognition

Textual emotion analysis is one of the three methodologies that this project employed on the development of a multimodal system that can diagnose mental health conditions. After further exploration, a suitable dataset that was found. This comprises a collection of anonymized conversations between patients and experienced psychologists, in which they address the patient's psychological state. The first column "context" contains the patient's concerns, the reasons and challenges that he's facing. Further the "emotion" and "response" columns represents the emotion of the patient displayed emotion and the doctor answer to the patient concerning's, respectively (cf.Table 6).

Table 6 – Text dataset data example

context	response	emotion
"I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm worthless and how I shouldn't be here.\n I've never tried or contemplated suicide. I've always wanted to fix my issues, but I never get around to it.\n How can I change my feeling of being worthless to everyone?"	"First thing I'd suggest is getting the sleep you need, or it will impact how you think and feel. I'd look at finding what is going well in your life and what you can be grateful for. I believe everyone has talents and wants to find their purpose in life. I think you can figure it out with some help."	sadness
"I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm worthless and how I shouldn't be here.\n I've never tried or contemplated suicide. I've always wanted to fix my issues, but I never get around to it.\n How can I change my feeling of being worthless to everyone?"	"I first want to let you know that you are not alone in your feelings and there is always someone there to help. You can always change your feelings and change your way of thinking by being open to trying to change. You can always make yourself available to learn new things or volunteering, give yourself a purpose."	sadness

Figure 6 illustrates the pipeline with employed data pre-processing steps to produce the final dataset (cf. Table 7) used to train and test the proposed model.

1. **Normalization:**
    a. Convert to String - Ensuring that the input text is in a consistent string format is essential, since data can sometimes come in different formats.
    b. Convert to Lowercase – Converting all characters to lower case ensures uniformity, which is important regarding the way that the models treat tokens.
2. **Remove Punctuation:** Punctuation marks (e.g., commas, periods, exclamation marks) do not usually contribute to the meaning of the text in most NLP tasks.
3. **Tokenization:** The splitting of text into individual tokens or words is fundamental to process text at the token level.
4. **Remove Stop Words:** Commonly words that carry no significant meaning in most contexts.
5. **Lemmatization:** Reducing the words to their base form proves to be important in the measure that the different forms of the same word are treated as a single entity.
6. **Remove Rare words:** Removing rare words that are often noise reduces the vocabulary size, improving computational efficiency.

Table 7 – Text dataset data example after preprocessing

context	response	emotion
“im going thing feeling barely sleep nothing think im worthless shouldnt ive never tried contemplated suicide ive always wanted fix issue never get around change feeling worthless everyone”	“First thing I'd suggest is getting the sleep you need, or it will impact how you think and feel. I'd look at finding what is going well in your life and what you can be grateful for. I believe everyone has talents and wants to find their purpose in life. I think you can figure it out with some help.”	sadness
“im going thing feeling barely sleep nothing think im worthless shouldnt ive never tried contemplated suicide ive always wanted fix issue never get around change feeling worthless everyone”	“I first want to let you know that you are not alone in your feelings and there is always someone there to help. You can always change your feelings and change your way of thinking by being open to trying to change. You can always make yourself available to learn new things or volunteering, give yourself a purpose.”	sadness

Finally, all that's left is to vectorize the text, since models cannot work directly with raw text and, afterwords, apply the data splitting for training and testing model.

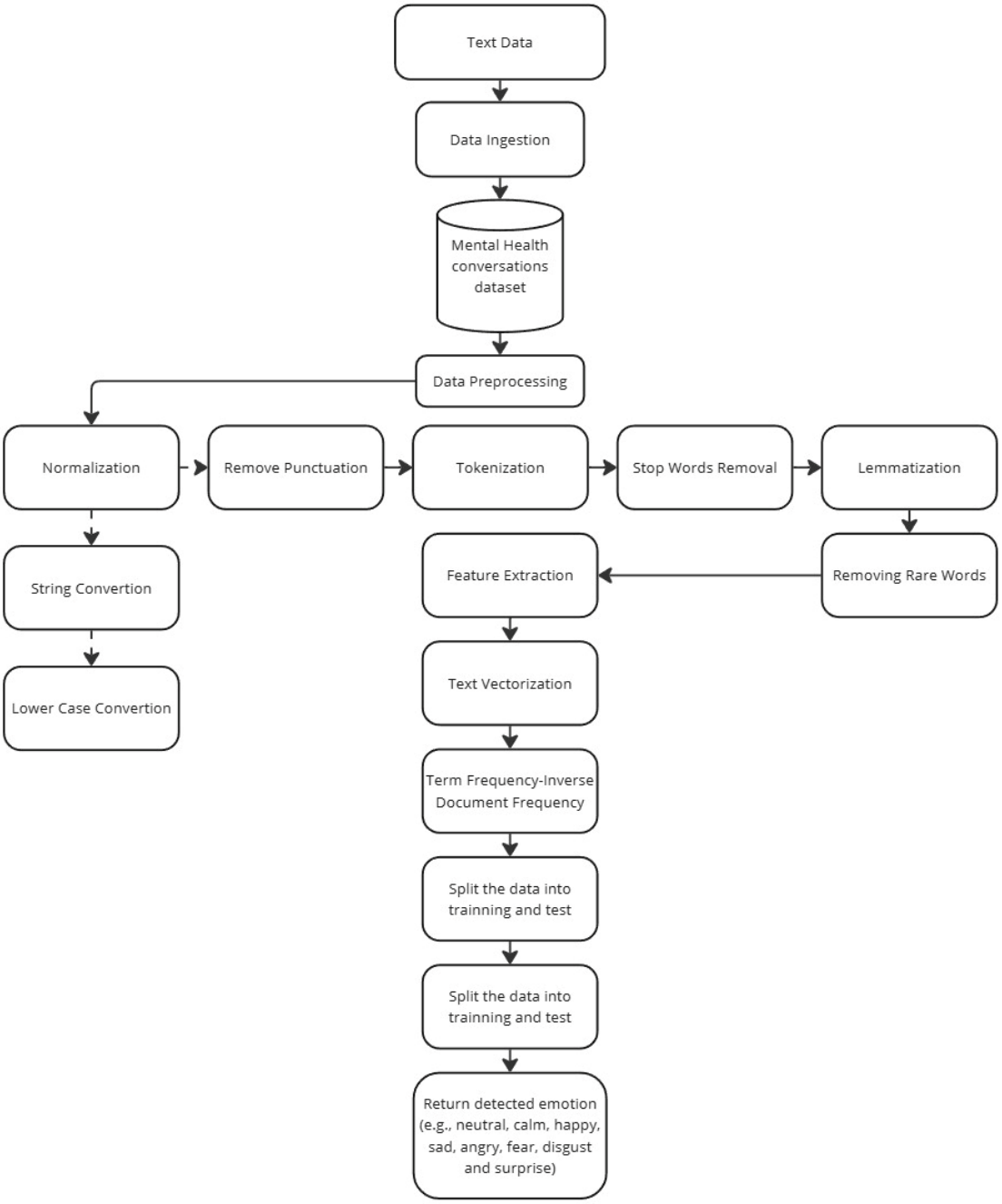

Figure 6 – Text Emotion Preprocessing Dataset Flow

3.3 Experimentation and Validation

The experimentation and validation phase of this project aimed to evaluate the performance of the proposed multimodal emotion recognition system. The evaluation process was designed to assess the system's robustness and generalization capability across the different modalities and datasets by evaluating metrics such as accuracy, precision, recall, and F1-score. Additionally, due to the studies observed on state-of-the-art, it was possible to set a base value for each of these metrics. Table 3 data benchmarks the varying performance levels across different modalities and datasets, emphasizing the importance of a robust multimodal approach. The proposed system aims to meet or exceed these scores by leveraging the strengths of each modality and ensuring balanced performance.

Furthermore, after training the individual models, the outputs of each modality are combined in a Json list and provided to an LLM along with the model metrics. The LLM serves as a fusion mechanism, interpreting the models' outputs and always having into account the performance characteristics of each modality to generate the final clinical report.

In addition, different augmentation techniques were also used to increase the size and diversity of the training dataset, aiming to improve the model performance by providing greater data variety:

Facial Emotion Analysis

To increase the diversity of the training data and improve the model's generalization capabilities, Keras' ImageDataGenerator class was employed for applying data augmentation techniques. Among the random series of transformations there were horizontal mirroring, zooming (reaching 20%), plus shifting along the width and height axes by 10% (cf. Figure 7). These alterations replicate variations that may appear in real-word contexts, aiding the model into growing more resilient to these changes.

Figure 7 – Application of Augmentation Application to Facial Data

Voice and Tone Emotion Recognition

To improve robustness and generalization for the speech emotion recognition model, feature extraction and then data augmentation techniques were applied to the audio signals. The purpose of these methodologies is to artificially expand the training dataset by introducing data variations, so they can mitigate overfitting and increase performance on novel information. Particularly, four enhancement techniques were used: additive noise, time-stretching, time-shifting, and pitch shifting. Supplemental noise appends minute adds Gaussian noise onto the audio signal since simulates ambient interference. Time-stretching modifies the length of the audio excerpt, keeping pitch unaffected so the model stays strong against variations in the speaking rate. Time-shifting rolls the audio waveform forward or backward by a random interval, ensuring that emotion recognition is not overly sensitive to the temporal start point. Pitch shifting adjusts the sonic frequency features within the audio. This action emulates natural undulations within vocal intonation (cf. Figure 8). Since this amplification procedure efficiently tripled the quantity of training data, it contributed substantially to the model's generalization capability across different speakers and recording conditions.

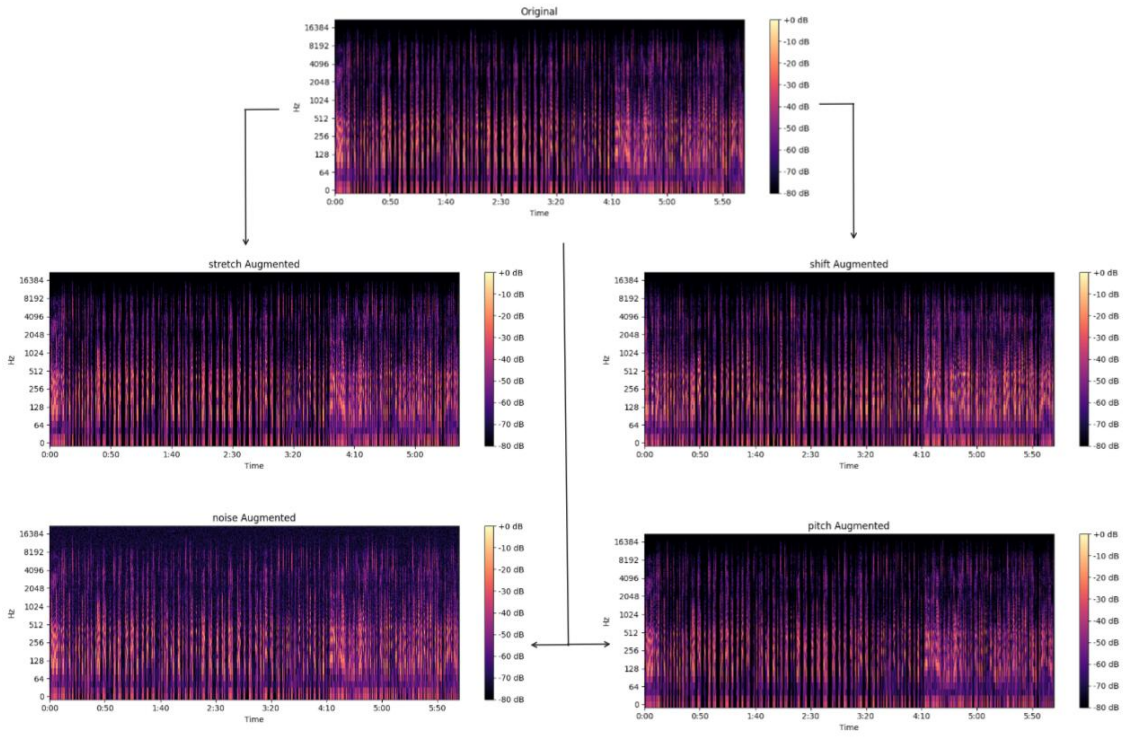

Figure 8 – Spectrograms of the Application of Augmentation Application to Voice Data

Sentiment Emotion Analysis

To enhance the robustness of the textual emotion recognition model, text augmentation techniques like paraphrasing and back translation were applied in order to improve the dataset diversity and reduce overfitting. The first one corresponds to the process where sentences are restructured while preserving their original meaning; the second introduces a natural variation by translating a sentence into another language and back into the original. Additionally, synonym replacement was also employed, which is responsible for replacing words with contextually appropriate alternatives to augment vocabulary diversity. Furthermore, random insertion and deletion of non-critical words were utilized to simulate natural changeability in human language expression (cf. Figure 9). The implementation of these strategies aims to improve the model's ability to generalize across different writing styles and phrasing patterns, strengthening its performance in real-world contexts.

Original: I feel anxious and overwhelmed by everything that is happening in my life.

Paraphrasing: I feel overwhelmed and anxious about everything that's happening in my life

Back Translation: I feel anxious and overwhelmed by everything that happens in my life.

Synonym Replacement: I smell anxious and overwhelmed aside everything that is happening in my life.

Random Insertion: I feel anxious happening and everything overwhelmed by everything that is happening in my life.

Random Deletion: I feel anxious overwhelmed by everything that happening in life.

Figure 9 – Application of Augmentation Application to Text Data

In another hand, fine-tuning techniques were also applied to the models:

Facial Emotion Analysis

Fine-tuning of the CNN architecture was carried out to improve facial emotion recognition performance on the FER-2013 dataset. The model underwent extensive tuning, incorporating five convolutional blocks with progressively increasing filter sizes, ranging from 32 to 512. To enhance generalization and reduce the risk of overfitting, batch normalization and dropout layers were integrated throughout the network. L2 regularization was applied specifically to the deeper convolutional layers to control model complexity. During training, Adam optimizer was employed with a low learning rate of 0.0001 to promote stable and gradual convergence. Furthermore, early stopping and model checkpointing mechanisms were used to automatically halt training when the validation accuracy stopped improving, ensuring that the best-performing model weights were retained.

Voice and Tone Emotion Recognition

To optimize the performance of the speech emotion recognition model, the training pipeline employed several fine-tuning strategies within. Adam optimizer was used to train the neural network. Categorical cross-entropy loss is indeed suited for use for classification problems having multiple classes. By dynamically adjusting the learning rate by monitoring the loss with a ReduceLROnPlateau callback, overfitting was prevented, and more effective model convergence was enabled during training. To reduce overfitting risk, regularization techniques such as dropout were systematically applied after the third convolutional block (20%) along with the first dense layer (30%). The model architecture involving four 1D convolutional layers, max-pooling, and dense layers was selected carefully to ensure balanced model complexity plus generalization ability. Input features were standardized through StandardScaler, and labels were one-hot encoded in order that the neural network had proper input formatting. These strategies stabilized the training process in total and made improvements to the model for generalization to unseen speech data.

Sentiment Emotion Analysis

The emotion classification model was implemented as a natural language processing pipeline using a RandomForestClassifier as the base model. The pipeline included tokenization, lemmatization, and contextual word representations to capture subtle emotional cues in patient–psychologist conversations. Hyperparameter tuning was conducted through a systematic search across parameters such as n_estimators and max_depth, combined with cross-validation to ensure robust performance and prevent overfitting. Optimizing these parameters allowed the model to achieve improved accuracy in detecting emotions expressed verbally, complementing insights from facial and vocal modalities.

3.4 Data Protection, Security and Ethics

The integration of Facial and Speech Emotion analysis into the proposed multimodal emotion analysis system implies a meticulous treatment of data protection, security, and ethical implications. This methodology involves the processing of biometric data, which is classified as sensitive personal data under the General Data Protection Regulation (GDPR)[48] and Artificial Intelligence Act (AI Act)[49].

3.4.1 General Data Protection Regulation

The General Data Protection Regulation is a European Union regulation that came into force in May 2018. The RGPD is a privacy and security law designed to give individuals more control over their personal data by defining fundamental rights. It outlines the obligations for the processed data, establishes methods for compliance and specifies sanctions for breaches. The regulation has the goal of strengthening individuals' rights

regarding their personal data and to ensure greater accountability and transparency from organizations that process such data. In this work context, which involves the utilization and the processing of the patient's data is essential to guarantee full compliance with the principles and requirements set out in the GDPR [48].

Lawfulness of Processing [48]

The context of the thesis and respective application involves the collection and processing of sensitive personal data – including biometric data (facial analysis), audio recordings and health related insights – which requires the establishment of a legal basis.

According to Article 6(1)(a), the ideal legal grounds for processing patient data is getting their clear and explicit consent. Additionally, since the system processes special categories of data, Article 9(2)(a) also applies, which requires explicit consent for processing special category data.

Purpose Limitation [48]

Under GDPR Article 5(1)(b), the personal data is required to be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes". The application must clearly define and present its purpose, which is:

- To analyze a patient's emotional state during therapy sessions.
- To generate therapeutic insights, reports and summaries to support the psychologist.
- To enhance the quality of mental health treatment.

The collected data must not be used for unrelated purposes, such as training models, commercial use, or third-party profiling, unless the patient is explicitly informed and permits it.

Data Minimization [48]

As per Article 5(1)(c) of the GDPR, data minimization requires only the collection of personal data that is adequate, relevant and limited to what is necessary for the stated purposes:

- Audio and video should only be recorded when necessary for the emotional and conversational analysis.
- If facial expressions are analyzed during the meeting, retain only model outputs, not raw frames, unless necessary for post-analysis.
- Post-session, segment the patient interactions only— do not retain psychologist speech for emotion analysis, used it only for context in summarization.

Accuracy [48]

In accordance with Article 5(1)(d) of the GDPR, the personal data must be "accurate and, where necessary, kept up to date". In this thesis context, this refers primarily to the outputs generated by AI models, such as Facial Emotion classifications, Sentiment analysis of speech and voice and Generated summaries and reports. Normally this principle only applies to factual personal data,

although, from an AI perspective, it also extends to inferred data – especially when such inferences may influence clinical decisions.

Storage Limitation [48]

As outlined in GDPR Article 5(1)(e), personal data must be "kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed." In this context, video recordings, transcripts, emotion inference results and at last the final session report, will be the impacted layers.

To be compliant with the GDPR, retention periods are defined for each data type in the app's data policy, which can be configured or extended by the psychologist if justified for ongoing treatment.

Finally, under the Article 17 – "Right to be Forgotten" – the patients can demand d deletion of their personal data before the retention period expires, unless exceptions apply (e.g., legal obligations or ongoing treatment needs).

Integrity & Confidentiality [48]

Following the GDPR's Articles 5(1)(f) and Article 32, it is required that personal data be "processed in a manner that ensures appropriate security," including protection against unauthorized access, accidental loss and destruction or damage. In terms of the AI pipeline this can be translated into:

- Encryption of data both at rest (stored outputs) and during transit (between server/client).
- Store audit logs of who accessed what and when.
- Apply access control mechanisms to ensure that only admins can view session data.
- Conduct regular security assessments, especially for third-party services (e.g., Google Speech to Text API, Ollama).

Accountability [48]

The GDPR (Article 5(2) and Article 24) mandates that not only should we be compliant with the GDPR, but we also must be able to demonstrate it. To do so both technical and documentation efforts are necessary:

- Document all processing activities, legal bases (e.g., consent), retention periods, and security measures.
- Maintain clear logs of when consent was collected, by whom, and for what purpose.
- Show that data subject rights are supported in the app (e.g., right to access, erasure).
- Maintain Data Processing Agreements (DPAs) with third parties (e.g. Google Speech to Text API, Ollama).

Data Protection by Design and by Default [48]

As stipulated in Article 25 of the GDPR, privacy and data protection are present into the system's design from the start—not added later. It should be assured that:

- Pseudonymization or anonymization techniques are utilized where possible.
- Ensure default settings are privacy-friendly (e.g., no recording unless consent is given).
- Limit access to only the psychologist involved in the session by role-based access control.

Data Protection Impact Assessment (DPIA)[48]

The DPIA becomes a mandatory requirement, which must be completed before the commencement of commercial processing, since the system fulfills multiple criteria for high-risk processing:

- Processing of Special Categories of Data on a Large Scale, the system manipulates sensitive data – for mental health diagnoses – and biometric data – facial and voice – on a potentially large user base.
- Systematic Monitoring and Evaluation, since the application involves systematic and extensive evaluation and scoring of individuals through automated methods to assess emotional states and generate therapeutic insights.
- Automated Decision-Making with Legal or Significant Effect, although the final diagnostic responsibility rests with the psychologist (Article 22), the AI outputs significantly influence clinical decisions, thereby having a substantial impact on the data subject.

The DPIA systematically identify, assess, and mitigate the high risks posed to the data subjects' rights and freedoms. The commercial entity must document the following:

- Risk Identification by conducting a detailed analysis of potential harms, such as bias in the AI model leading to incorrect diagnoses, unauthorized access to sensitive health records, or re-identification of pseudonymized data.
- Mitigation Measures are necessary and a clear plan for reducing the identified risks is required. This includes implementing robust measures beyond basic encryption, such as Advanced Pseudonymization Techniques, enhanced Access Control specific to clinical roles, and a strict Audit Log system to track all data access and processing activities.
- Formal documentation within the DPIA confirms that, despite the AI's role, human intervention is a required step to ensure that the final decision is not solely automated (Article 22).

Finally, if the DPIA concludes that, even after all envisaged mitigation measures are applied, the residual risk to data subjects remains high, the entity is legally required to consult the relevant Supervisory Authority (Article 36) prior to initiating the processing. This ensures that the

authority reviews the system's compliance and risk management before the product is placed on the market.

Records of Processing Activities [48]

While the entirety of Article 30 obligations typically applies to organizations with more than 250 employees or those doing high-risk processing, this application must still maintain Records of Processing Activities since it records and processes special category data (cf. Table 8).

Table 8 – Record of Processing Activities for the Emotion Mental Health Detection

Processing Activity	**Purpose**	**Legal Basis**	**Data Types**	**Retention**	**Recipients**
Record video session	Emotional & conversational analysis	Consent	Audio, video, face, voice	Defined by the system admin	Internal (psychologist)
Speech-to-text (Google API)	Transcription for analysis	Consent + DPA	Audio → text	Until report	Google (processor)
Emotion classification (3 models)	Patient emotion recognition	Consent	Facial features, speech tone, text	Until report	Internal only
LLM summarization (Ollama)	Generate report	Consent + DPA	Processed text, emotion labels	Until report	LLM host (processor)
Final report storage	Support psychological care	Legitimate interest	Summary, insights	Patient-controlled	Psychologist dashboard

Processor Obligations [48]

Article 5(1)(c) of the GDPR states that a processor is any third party that processes personal data on your behalf, in this case Google Speech-to-Text API and Ollama. Following that, the application the administrator is required to:

- Ensure a written DPA is in place with all processors (Art. 28(3)).
- Only use processors that provide sufficient guarantees regarding GDPR compliance.
- Ensure processors process data only under documented instructions (e.g., no reuse for training).

Data Subject Rights [48]

Pursuant to Articles 12–22 of the GDPR, the patients should be able to exercise the following rights:

Table 9 – Patient's Rights

Right	Article	Description	Implementation
Access	Art. 15	Know what data is held and how it's used	In-app portal or email request for session data, summaries, and processing logs
Rectification	Art. 16	Request correction of inaccurate data	Allow psychologists to edit or override incorrect AI outputs in reports
Erasure	Art. 17	"Right to be forgotten"	Option to delete raw session data, transcripts, or reports upon request
Restriction	Art. 18	Temporarily halt processing while a request is reviewed	Disable further analysis or report generation if a dispute arises
Data Portability	Art. 20	Receive data in a machine-readable format	Export transcripts, emotion labels, and summaries in PDF/JSON format
Objection	Art. 21	Object to processing based on legitimate interest	Allow opt-out from analytics features (if not purely based on consent)
Automated Decision-Making	Art. 22	Protection from decisions made solely by algorithms	Ensure psychologist reviews and validates all reports before acting on them

Automated Decision Making & Profiling [48]

Under GDPR (Article 22), data subjects have the right not to be subject to a decision based solely on automated processing, including profiling, if it significantly affects them (e.g., legal, professional, psychological impact). This means that this application should only be used as a tool to provide insights, notes and supplementary analysis to the psychologist. This system is not responsible per automated decision-making, that responsibility still resides within the phycologist – human validation.

Consent Requirements [48]

Following the GDPR's Articles 6, 7, and 9, since the current system utilizes special category data, explicit consent is the primary lawful basis. The consent should be:

- **Freely given**: No coercion, the use of the app must be optional.
- **Specific**: Separate consents for Recording the session, Emotion/sentiment analysis and use of LLM summarization.
- **Informed**: Provide clear explanations of what data is collected, how it will be used, and the risks.
- **Unambiguous**: Do not use pre-checked boxes to ask for consent.
- **Documented**: Register when and how consent was given.

The application also includes a clear and accessible mechanism to withdraw consent at any time, as required by Article 7(3).

3.4.2 AI Act

The Artificial Intelligence Act is a proposed regulation from the European Union. It has the aim of making sure that AI systems are safe and clear, while also respecting fundamental rights. The act establishes a risk-based framework that classifies AI based on the level of risk they pose, requiring stricter rules for high-risk systems. It is crucial for AI applications that handle sensitive data or have major effects on society, like those used in healthcare, in which this work is integrated, to comply with the AI Act.

Risk Management System

Under the AI Act Article 9, protocols state that high-risk systems must implement a continuous and iterative risk management system throughout the entire lifecycle of the AI. To achieve this is necessary:

- **Identify and categorize risks:**
  - Misclassification of emotions (false positives/negatives).
  - Overreliance on AI-generated summaries.
  - Potential psychological harm due to biased or inaccurate outputs.
- **Analyze and estimate the severity of each risk.**
- **Design mitigation strategies**, such as:
  - Psychologists must always verify the outputs correctness.
  - Setting thresholds for AI confidence before displaying results.
  - Fallbacks for model failure (e.g., set default emotion tag "No emotion detected").
- **Test and evaluate** risks during development (e.g., stress tests, error injection).
- **Document and monitor** these risks of post-deployment.

Data and Data Governance

Since the system's purpose is to analyze emotions through the training and usage of models it is covered by data governance requirements:

- **Training, validation, and test data are:**
  - Relevant and representative (avoid biases based on race, gender, etc.) which is ensured since the system uses multi-gender and multi-ethnic datasets.
  - Free of errors, and clearly labeled (e.g., emotion tags must be well-defined).
  - Preprocessed in a traceable manner, with documentation of transformations (e.g., face normalization, noise filtering).
- **No discriminatory outcomes**: models should not disproportionately mislabel specific demographics.
- **Keep records of:**
  - Data source, licensing, and purpose.
  - Annotation methodology and annotator qualifications (if applicable).
  - How data was cleaned, augmented, or filtered.

Technical Documentation

As stated in Article 11 of the AI Act, technical documentation and artifacts should be maintained as regulatory layers to evaluate the system safety and compliance:

- **System Overview:** represented in workflows, pipelines, etc.
- **Model documentation**: for each AI model (facial, vocal, textual sentiment, LLM):
  - Architecture, training methods, datasets used.
  - Accuracy metrics, confidence levels, known limitations.
  - Versioning and update policy.
- **Data management** procedures (Art. 10).
- **Risk management** outcomes (Art. 9).
- **Human oversight mechanisms**.
- **Performance metrics and validation tests**: including robustness to edge cases and adversarial inputs.
- **Security controls**: encryption, access management, anti-tampering.

Record-Keeping

Under AI Act rules (Article 12), high-risk AI systems must generate and store logs to trace operations throughout their lifecycle. This ensures that, in case of an incident, error, or audit, it's possible to trace back what happened and why.

Transparency and Information to Users

According to Article 13 AI Act the users of the application (psychologist and the patient) must be clearly informed that they are interacting with an AI system and understand how it functions

at a high level. Specify the stages of the AI process (e.g., transcription, emotion analysis, report generation), detail the purpose of each model, describe the system limitations and finally expected performance and reliability – The observations provided are grounded in technical and empirical evaluation, but do not guarantee 100% reliability, which is why is clearly labeled "*This is an AI-generated summary*".

Human Oversight

Article 14 of AI act asserts that to prevent misleading information and ensure that AI errors do not directly influence decisions without review, the final report should be human reviewed. In the thesis system, the psychologist will adopt this role and confirm, discard or override AI results.

Accuracy, Robustness, and Cybersecurity

As outlined in AI Act Article 15, AI system must achieve and maintain a high level of accuracy, robustness, and cybersecurity, since it directly influences human psychological assessment.

The accuracy should be reported for each AI component – facial emotion recognition, voice tone model, text sentiment. Also, confidence thresholds can be placed for acceptable predictions. To access the robustness of the system, test can be made under variable conditions to evaluate this metrics – stress testing. Finally, following the best principles of cybersecurity the data must be encrypted (in transit and at rest), access and authentication controls must be implemented and logs enabled monitoring these protocols.

Post-Market Monitoring

In accordance with Article 61, even after deployment it is necessary to maintain mechanisms to monitor the AI system behavior, to assess performance, risk or compliance failures. To ensure the system stays aligned with real-world needs, psychologists' input must be continuously gathered—whether they've noticed issues with accuracy, reliability, unexpected outputs or even ethical concerns.

Conformity Assessment

As outlined in Articles 43–51 of the AI Act, high-risk systems must undergo through a conformity assessment before being launched to the market. This means that:

- All technical documentation, risk management, data governance, and human oversight should be completed and certified.
- Every component should be tested and validated.
- Compliance via a declaration of conformity is guaranteed.

Provider Obligations and Market Access

While the previous sections detail the mechanisms of compliance, the application commercial transition necessitates fulfilling the formal obligations of the Provider before and after launching the high-risk AI system on the market. The Provider, as the responsible entity, must ensure continuous and demonstrable compliance through the commercialization process, by securing key obligations:

- Before being launched on the market, the high-risk AI system must be formally registered in the EU-wide public database managed by the Commission (Article 51).
- Establishing, documenting, and maintaining a robust Quality Management System to ensure adherence to all AI Act requirements (Article 17). The Quality Management System must cover all stages from design, development, production, deployment, to post-market monitoring.
- Retaining all technical documentation and the results of the conformity assessment for a period of 10 years after the system is placed on the market (Article 18).
- Implementing corrective actions immediately if the system is deemed non-compliant or poses a high risk to fundamental rights after its launch.

Furthermore, the commercial Provider to comply with the general monitoring framework, must obey the following documentation duties:

- The Provider must establish a formalized system to record and report any serious incidents (e.g., an error leading to patient harm or a violation of fundamental rights) to the relevant national authorities immediately after becoming aware of them. This ensures that market surveillance bodies are promptly informed of any performance or safety failures.
- The Provider must supply clear and concise instructions for use to professional users (Article 13), specifying:
    - The precise purpose and intended use of the system clearly defines the scope of the AI's diagnostic support.
    - The known limitations and potential biases of the model, advising the professional on scenarios where the AI's output may be less reliable (e.g., specific demographic groups, complex emotional states).
    - The type of expertise and training required for the human user (the psychologist) to correctly interpret the AI's output, thereby reinforcing the mandatory nature of human oversight.

Summarizing, the Provider responsibilities ensure that the legal and ethical accountability for the safety and reliability of a high-risk AI system rests firmly with the entity placing it into commercial operation.

3.5 Discussion

This chapter outlined the elements used to evaluate along with the construction of the multimodal emotion recognition system proposed. It begins with an introduction to the tools, methodologies, and datasets selected across three modalities – facial expression, speech and voice, and textual emotion analysis. Each dataset was carefully selected based on its representativeness, diversity, and alignment with standardized emotional categories.

The chapter delineated the preprocessing steps that were used on each data type for improved consistency as well as model performance— ranging from image normalization and data augmentation in facial emotion recognition, to advanced audio signal transformations for speech analysis, and NLP preprocessing techniques for text data. The procedures confirm the data input remains pristine, organized, and furthermore resilient to actual fluctuation.

Experimentation and validation processes were explained through the employment of individual model evaluation metrics (accuracy, precision, recall, and F1-score), model fusion strategies, and fine-tuning approaches tailored for each modality. The model's resilience was additionally elevated via integrating enhancement methodologies – data augmentation.

Finally, the chapter addressed critical ethical, legal, and data protection considerations. It highlighted compliance with GDPR and the AI Act, ensuring lawful data handling, privacy by design, risk management, and human oversight mechanisms. The compliance with these two regulations was not taken lightly on the contrary, specifically the need for Explicit Consent (GDPR Article 9), ensure that the session recording only starts after the patient’s digital consent is actively signed. Furthermore, since the system is classified as High-Risk under the AI Act, a continuous Risk Management System (AI Act Art. 9) is maintained throughout its lifecycle, informing the requirements for Technical Transparency in Reporting and the established Human Oversight methodology (AI Act Art. 14). This oversight obligates the psychologist to always validate, confirm and if necessary, override the AI generated reports before taking any clinical action (GDPR Art. 22), thereby converting the solution from a prototype into a legally responsible diagnostic support tool.

4 Implementation, Validation and Results Discussion

This chapter outlines the process of moving from the conceptual design phase to the implementation of a fully functional multimodal emotion recognition application, which will serve as a tool and an asset for professionals in the mental health sector.

It starts by introducing the system components, the front-end and back-end, describing their implementations, their roles, and how they interact with each other, presenting a view of the application as a whole. Furthermore, this chapter mentions and explains the two main flows of the system: the real-time and post-session workflows.

Sequentially, the performed validation procedure is explained, which is responsible for validating and attributing a trust level to the developed system. This step introduces and details the usage of a set of standardized self-rated screening measures - the DSM-5 Self-Rated Level 1 Cross-Cutting Symptom Measure –, applied to guide the application validation process and assess the system's practical applicability.

Finally, the validation results are presented and discussed, emphasizing the domains specified by the DSM-5, with particular attention to areas where the system demonstrates strong predictive performance, while also addressing its limitations in detecting subtler domains. The results expose and assess global performance metrics, including accuracy, precision, recall, F1-score, and Cohen’s Kappa, providing a comprehensive view of the system’s reliability.

4.1 System Overview and Architecture

The proposed AI-driven emotion recognition system operates through a modular client–server architecture which contains two essential parts: a back-end processing engine and a front-end interface. The system design provides scalability alongside maintainability features and future model and functionality integration capabilities (cf. Figure 10).

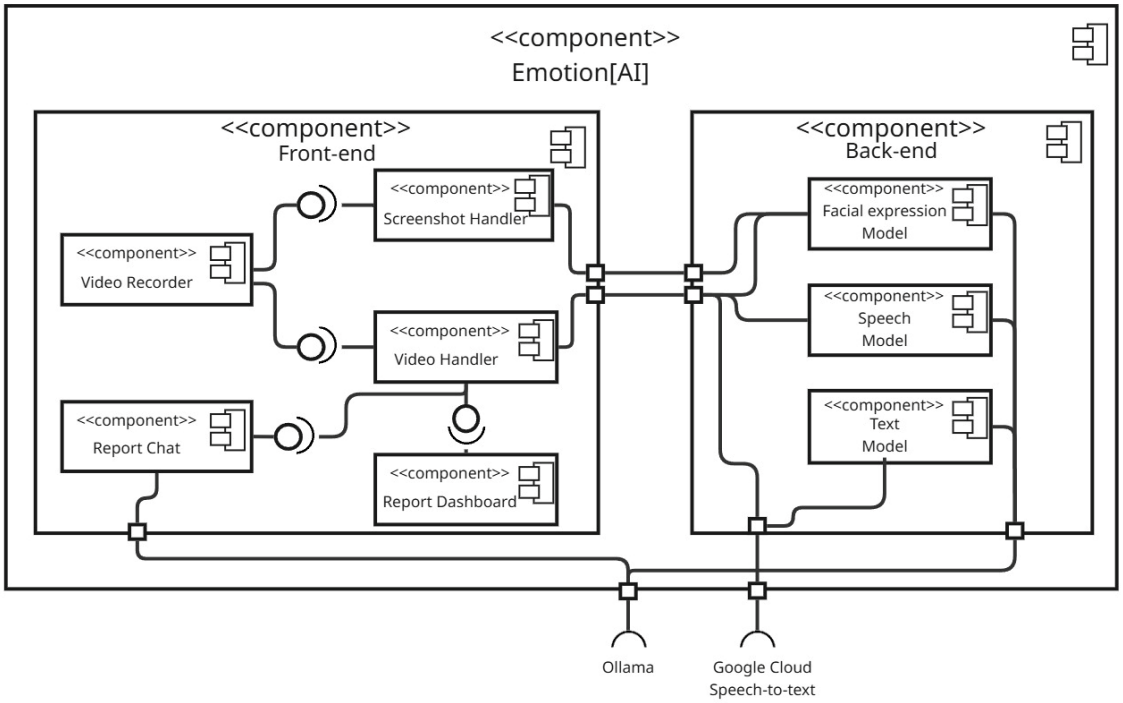

Figure 10 – Logic View Level 3 (View Model 4+1 [50])

4.1.1 Front-End Architecture

Healthcare professionals interact with the system using a front-end interface developed utilizing Angular and TypeScript (cf. Figure 11). The interface is designed to be user-friendly, intuitive, and responsive, allowing the users to use it across different devices. Through the interface the therapists will be able to conduct the meetings and access all the reports and data generated from previous consultations. The architecture design in angular, is implemented to ensure scalability and maintainability, which means that is prepared to receive future enhancements or additional features and integrations.

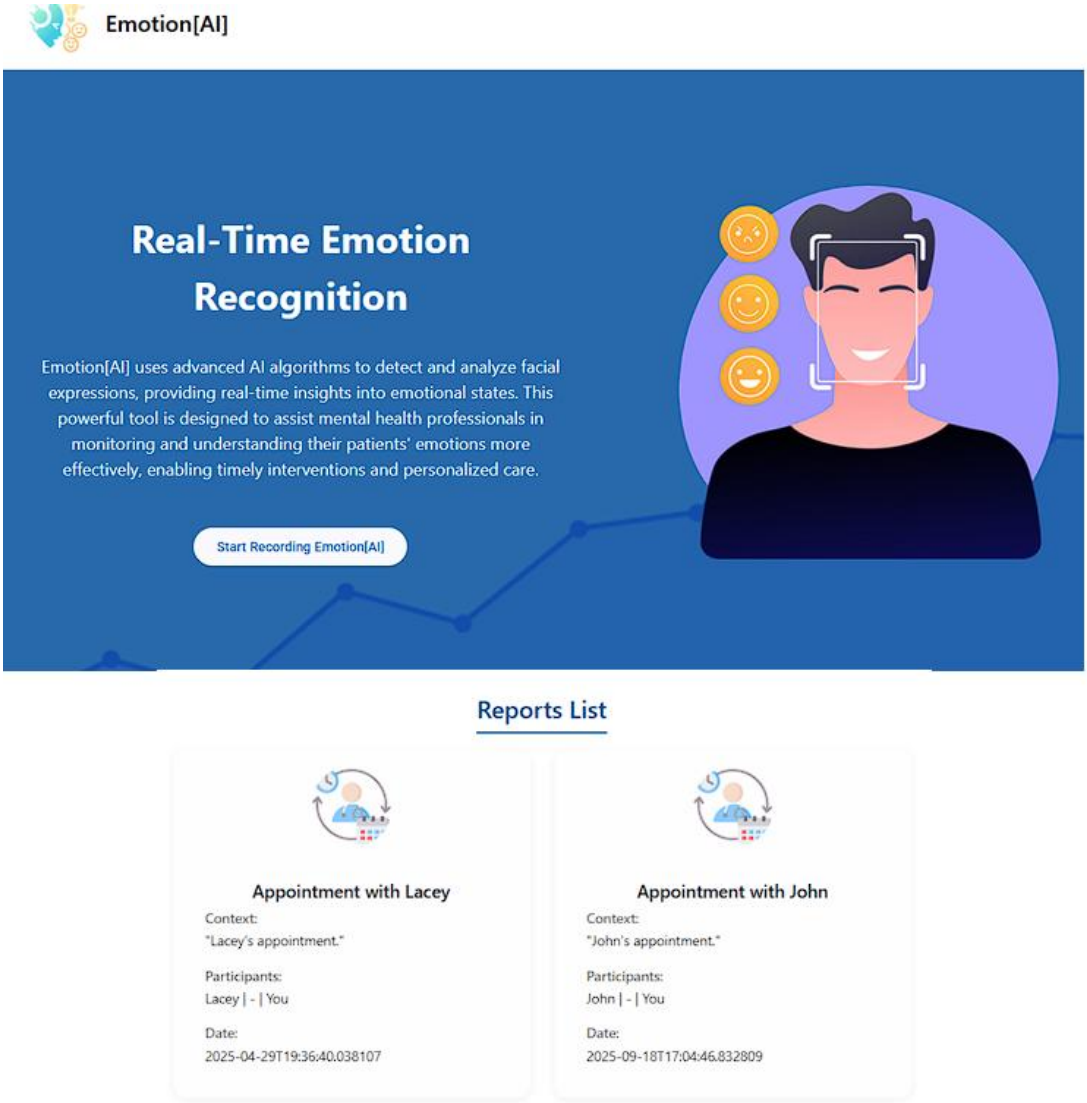

Figure 11 – Application Homepage

Through this interface healthcare professionals can initiate video recordings of patient-therapist sessions which get transmitted to the back-end processing system. The interface provides two separate operational functions:

Real-Time: Captures video stream screenshots at regular 5-second intervals during active video transmission. The back-end facial expression recognition model receives these images directly to identify patient emotions in real-time. The system shows real-time results which provide therapists with immediate feedback during their sessions. Additionally, a scatter chart was added and is responsible for keeping track of the emotions felt by the patient during the therapy session, which helps therapists to visualize the patient's emotional progression through time, making it easier to identify patterns or triggers, and create their interventions accordingly (cf. Figure 12).

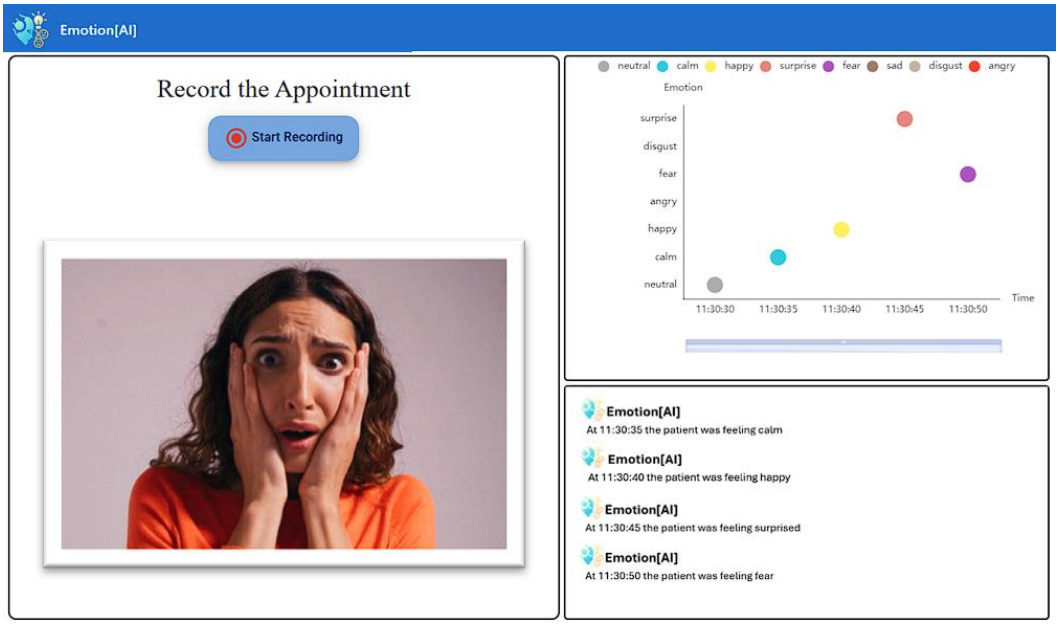

Figure 12 – Application Recording Page

Post-Session: At the end of the session, the recorded meeting is saved and uploaded to the back-end for comprehensive processing. After the back-end finishes processing the data, a new card appears on the report list, located at the bottom of the application homepage. Each card corresponds to a patient session, and by clicking on it, the front-end redirects to a new page that contains the report for the respective session, including the patient's data and meeting details. Furthermore, a chatbot was also implemented, which allows the psychologist to interact with the report data conversationally. The chat is prepared to answer specific questions about the session, such as summarizing the patient's most displayed emotions alongside the questions that triggered them. It can highlight moments of emotional shift, providing context and possible reasons for it. On the bottom line, since it has access to all the historical data sessions, it can track progress towards therapeutic objectives and flag recurring patterns that might require specific clinical attention (cf. Figure 13).

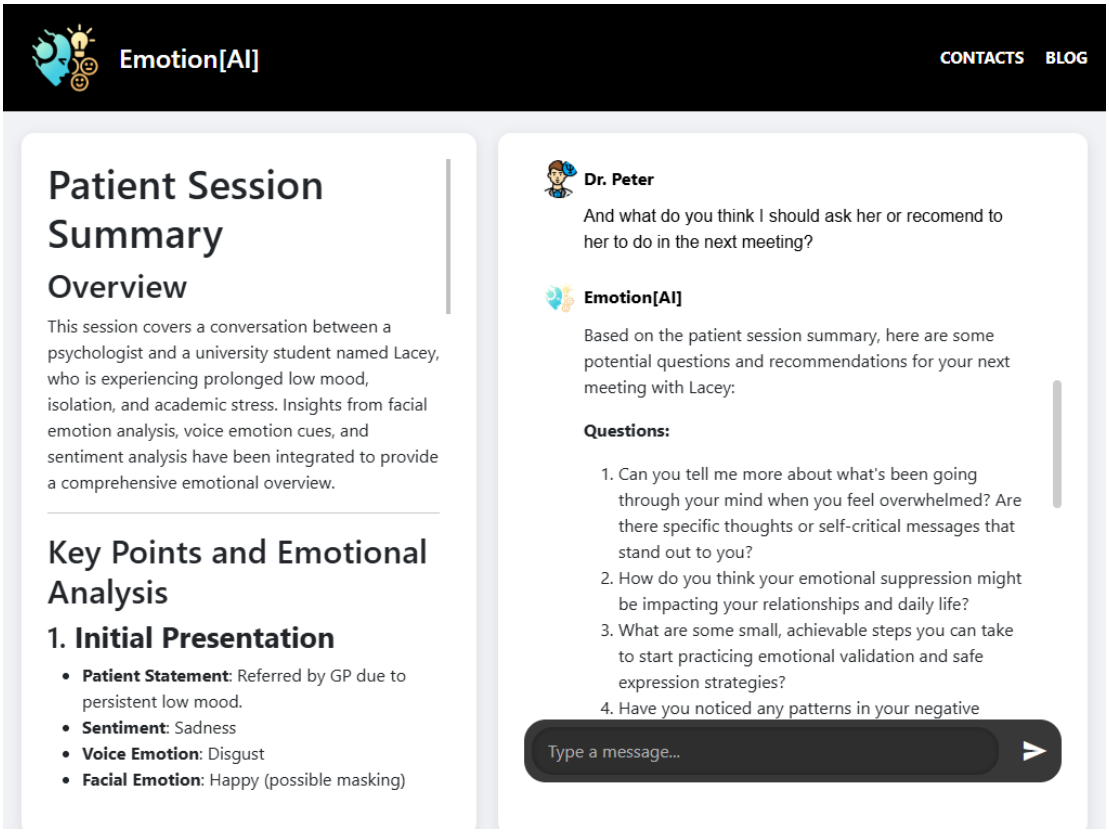

Figure 13 – Application Report and Chatbot Page

4.1.2 Back-End Architecture

The back-end starts its data handling once it receives the data from the front-end. Since the application is developed to manage two different workflows – real-time facial emotion analysis and post-session report generation – it receives two different types of data depending on the current operational mode.

For the real-time analysis workflow, as described previously, the front-end captures images at regular intervals that are afterward sent to the back-end to be processed. The system accepts the image input as an encoded Base64 string, which is received through a JSON payload. The input starts being decoded into a raw image format and sequentially converted into a grayscale color space to match the model's training data input specifications (FER-2013). It is then resized and normalized before being transformed into a tensor representation for the model input. Before being fed to the facial emotions model, the input goes through a pre-trained Multi-task Cascade Convolutional Neural Network, which is configured for optimal face detection with parameters such as a minimum face size of 200 pixels and set confidence thresholds. This model scans the image and returns the coordinates x, y, width, and height of a bounding box containing the most promising human face detected. Next, the region of interest determined by the bounding box is cropped from the original image, isolating the identified face that will be prepared, as before, to ensure compatibility with the emotion CNN model. Subsequently,

the preprocessed image is fed to the fine-tuned emotion CNN model, which consists of five core convolutional blocks with progressively increasing filter sizes (32 to 512). Each of the blocks is improved with batch normalization and dropout layers to enhance the generalization and reduce overfitting. In addition, L2 model regularization is applied to the deeper levels of the network to control its complexity. Regarding the model's output, the network performs a forward pass and converts its results into a probability distribution across the seven available emotion types: angry, disgust, fear, happy, sad, surprise, and neutral. The emotion label with the highest computed probability is selected as the final prediction for the input image, which is converted into a JSON object and returned to the front-end that is responsible for displaying it on the Application Recording Page (cf. Figure 12, right side).

For the post-session report generation workflow, it starts once a session is concluded and corresponds to the moment when the system begins a processing workflow to generate the meeting's detailed report. This post-session pipeline is different from the real-time analysis; it consists of a multi-modal analysis that processes the full meeting data, sent from the front-end, in three emotion detector models - Emotion Facial analysis, Emotion Voice Recognition, and Emotion Text analysis - to provide the therapist with a deep, retrospective evaluation. Therefore, upon the session termination, the front-end application aggregates the collected data - time timeline of emotions captured in real-time, session metadata (e.g., patient ID, therapist ID, date, duration), and the recorded video file— into a data packet, and transmits it to the back-end.

The first stage of the pipeline involves extracting and preprocessing the raw video data, starting by extracting its audio content and producing a transcript of the speaker's speeches annotated with their information. The extracted audio is converted to a mono-channel WAV format with a sampling rate of 16 kHz, ensuring compatibility with the Google Speech-to-Text API, which will be responsible for producing the transcript. The resulting audio file needs to be uploaded to a Google Cloud Storage bucket, since the transcription API operates on files saved in this storage unit. Once uploaded, it is ready to be consumed by the API, which is configured with speaker diarization enabled and a predefined number of two speakers. The API returns the full transcription, including for each individual intervention the associated speaker tag (Doctor or Patient) and the respective time boundaries (cf. Figure 14).

```
{
    "speaker": "Doctor",
    "message": "Hi Lacey. Nice to meet you. So, I understand that you referred here by your GP
because you've been feeling quite down recently.",
    "start_time": 19.5,
    "end_time": 28.2
},
{
    "speaker": "Patient",
    "message": "Yeah, I've been feeling quite bad for quite a while, so I thought maybe it's t
ime to see someone about it because I don't want to feel like this anymore.",
    "start_time": 28.4,
    "end_time": 37.5
}...
```


Figure 14 – Example of Audio Transcription

Following the generation of the transcript, the next stage of the post-session workflow focuses on the facial analysis of each patient's intervention. Beginning by using OpenCV, the recorded video is read, and an extraction of all frames within the given time interval is conducted (for each intervention, the transcript has the *start_time* and *end_time* parameters). This step isolates the visual data corresponding exactly to each patient's speech segment. Each of the extracted frames will follow the same process described on the real-time workflow, from the face detection on the pre-trained Multi-task Cascade Convolutional Neural Network to the fine-tuned emotion recognition CNN, which will give the final emotion output.

The next stage consists of using the time-aligned transcript to isolate the patient's interventions and identify the dominant emotions expressed in the acoustic patterns of his voice. This step relies on a beforehand trained and ready-to-use model, developed using several public emotional speech datasets (such as CREMA-D, RAVDESS, SAVEE, and TESS), and preprocessed as described in Chapter 3, section 3.2.1**.** The trained model, along with the label encoder and feature scaler (both also saved when the model was trained), is loaded at runtime to perform inference on the data, without the need to retrain. To ensure that only the patient's speech is analyzed, given the start and end times of each patient's utterance, the system can extract the corresponding audio portion from the full recorded session. The extracted audio segments are, afterward, converted into a numerical feature vector (using the feature-extraction techniques described in Chapter 3, section 3.2.2), capturing both spectral and temporal aspects of the speech signal that are relevant to emotion recognition. The saved scaler object is then applied to normalize the feature vector, ensuring that it is consistent with the data distribution the model was trained on. The normalized feature vector is fed into the pre-trained model (one-dimensional convolutional neural network), which has as output a probability distribution over the available emotion classes (angry, disgust, fear, happy, sad, surprise, and neutral). The label encoder is at last used to translate the predicted class index back into its corresponding emotion label.

The fourth stage of the post-session pipeline is responsible for analyzing the emotional content conveyed in the text of the patient's utterances. While the vocal and facial steps assess how the

message is delivered, this module focuses on the message itself, in what is said by the patient. Natural Language Processing is the technique used to detect intrinsic emotions in the transcribed speech. This module relies on a trained and previously saved model (a Random Forest Classifier), together with its corresponding Term Frequency–Inverse Document Frequency (TF-IDF) vectorizer, both stored locally during the model training. Each transcribed segment is preprocessed through a band of text-cleaning and normalization (as described on Chapter 3, section 3.2.3). The cleaned text is then transformed into a numerical feature vector using the saved TF-IDF vectorizer to ensure that the vocabulary and weights are consistent with those used during training. The resulting vector is then passed to the Random Forest Classifier, which outputs the most strongly expressed emotion in the text, among the available types (angry, disgust, fear, happy, sad, surprise, and neutral).

Upon the completion of the three emotion models, it produces a structured representation of the session, in which the patient's interventions are organized along with the corresponding multimodal information. Additionally, this list of dictionaries also contains for each utterance the start and end times and the transcribed message (cf. Figure 15).


```
  {
    "speaker": "patient",
    "message": "I just don't see the point anymore. I try to get out of bed, but everything fe
els so heavy and hopeless.",
    "start_time": 125.4,
    "end_time": 132.1,
    "message_sentiment": "sad",
    "voice_emotion": "fear",
    "facial_emotion": "sad"
  },
  {
    "speaker": "patient",
    "message": "No, I'm fine, really. It's all fine. Let's just move on.",
    "start_time": 218.7,
    "end_time": 222.9,
    "message_sentiment": "neutral",
    "voice_emotion": "fear",
    "facial_emotion": "sad"
  }
```


Figure 15 – Example of a post-session multimodal transcript entry

In addition to this post-session data, the real-time emotion previously collected data is gathered along with the full transcript, the model's metrics (accuracy, recall, precision, and f1-score), professional psychological reports, and therapeutic techniques, to ensure that the final report is clinically informed and contextually accurate. All this information, together with a structured prompt that instructs the LLM on the desired format (cf. Figure 16), headings, and level of clinical detail, is then sent to an Ollama LLM, which will be responsible for analyzing it, and finally creating a cohesive post-session report.

```
# Instructions for Generating the Session Report

**Objective:** Create a comprehensive and clinically useful summary of the patient's therapeut
ic session by analyzing multimodal data.

## Core Requirements

*   **Summarize** the **emotional progression of the session**, highlighting changes in patien
t affect over time.
*   **Identify** key patterns or triggers in patient behavior, speech, and emotions.
*   **Cross-reference** multimodal data: text sentiment, voice emotion, and facial emotion.
*   **Provide** clinically informed **observations and insights**, grounded in psychological p
rinciples and techniques.
*   **Suggest** actionable recommendations for follow-up sessions, interventions, or coping st
rategies.

## Report Structure

Structure the report clearly using the following headings:

1. **Key Points and Emotional Analysis**
2. **Session Overview**
3. **Patient Emotional Analysis**
4. **Therapist Observations**
5. **Key Interventions and Reactions**
6. **Recommendations and Next Steps**

## Tone and Style

Maintain professional, objective, and empathetic language throughout.

## Output Format

Provide the final report in clear, structured text, suitable for the therapist to read directl
y.
```

Figure 16 – Prompt provided to the LLM to guide the generation of the report

In another hand, is important to underline that when the emotion recognition models produce conflicting outputs (e.g. facial expression shows sadness while speech tone shows anger), the LLM is responsible per analyzing the discrepancies by considering context from all modalities, the session conversation, model's metrics, and patterns of emotional expression which are present in the provided professional reports. It reasons why the models' results may differ; for example, subtle speech cues may indicate a hidden emotion not visible in facial expressions, or temporary facial expressions may not reflect what the speaker is saying. Based on this reasoning, the LLM reconciles the differences to infer the most plausible emotional state, producing a more accurate and interpretable assessment for the report.

To conclude, all received and generated data, along with the post-session reports, are persisted in a PostgreSQL database. This ensures that all session information is securely stored and can be accessed for future reference. From the front-end, on the Application Report Chatbot page

(cf. Figure 13), the therapist can interact with the stored data through the chatbot and ask it everything he needs from the patient-therapist meeting.

4.2 System Validation and Results

The validation of the diagnosis support tool is a fundamental step to ensure its reliability and effectiveness in real-world applications. To guarantee the scientific credibility of the proposed system, a validated clinical instrument – DSM-5 Self-Rated Level 1 Cross-Cutting Symptom Measure – was used as a standard reference for comparison and system trustworthiness assessment.

The selection of this scale ensures the system validation by applying a structured methodology that aggregates subjective (self-assessment) and objective indicators (analysis of facial expressions, vocal patterns, and verbal content). The following subsections will focus on experimental validation strategy, which utilize instruments designed to establish a relationship between the system's outputs and DSM-5 indicators, as well as the evaluation metrics adopted and applied to analyze the agreement between the two methods.

4.2.1 Methodology

The validation strategy is based on the comparison of the results generated by the system and the collected data using the DSM-5 Self-Rated Level 1 Cross-Cutting Symptom Measure. The DSM-5 scale is composed of 23 questions that evaluate the symptoms of 13 clinical domains, including: Depression, Anger, Mania, Anxiety, Somatic symptoms, Suicidal ideation, Psychosis, Sleep disturbances, Memory problems, Repetitive behaviors, Dissociation, Personality functioning, and Substance use. Each item is classified on a scale from 0 to 4 (from "None" to "Severe/Almost every day"). Scores higher than 2 indicate the necessity for further clinical evaluation.

Table 10 – DSM-5 Symptom Screening Tool: A 23-Item Multidimensional Assessment

Domain	**Questions (Have you been …)**
Depression	Feeling pleasure or interest in doing things?
	Feeling down, depressed, or hopeless?
Anger	Feeling irritated, moody or angry than normal?
Mania	Sleeping less than usual but still have big reserve of energy?
	Taking more risks than usual?
Anxiety	Feeling nervous, anxious, scared, worried, or tense?
	Having panic attacks or intense fear?
	Avoiding situations due to anxiety?
Somatic Symptoms	Experiencing unexplained pains (e.g., head, back, abdomen)?
	Feeling that your health concerns were not taken seriously?
Suicidal Ideation	Thinking about harming yourself or ending your life?
Psychosis	Hearing voices when no one was around?
	Feeling that someone could hear your thoughts, or you could hear others' thoughts?
Sleep	Having sleep problems that affected your rest quality?
Memory	Having memory or orientation issues (e.g., forgetting familiar routes)?
Obsessions & Compulsions	Getting unwanted thoughts or images repeatedly invade your mind?
	Feeling the need to repeat actions or thoughts?
Dissociation	Experiencing detached periods from your body, surroundings, or memories?
Personality Functioning	Feeling unsure about who you are or what you want in life?
	Feeling distant or disconnected from others?
Substance Use	Drinking 4 or more alcoholic drinks in a single day?
	Using tobacco (cigarettes, cigars, etc.)?
	Using prescription drugs without a prescription or in excess?

This approach allows to verify at which level the emotions captured through facial, vocal, and text analysis are compatible with the symptoms reported directly by the patient (self-report). Having that in mind, a set of DSM-5 clinical domains (e.g., depression, anxiety, anger, suicidal

ideation, among others) was defined, and a bridge between these domains and system data was set.

4.2.2 Cross-Validation with DSM-5 Diagnostic Standards

Mapping the different types of data processed by the system with 13 clinical domains assessed by level 1 of the DMS-5 was the next step to establish a valid comparison between the results of both. This representation allows make an interpretation of whether the captured signals – facial expressions, voice, and text – may be related to clinical psychological assessment, as described in the DMS-5.

Facial expressions are often the most telling when it comes to picking up on strong, immediate emotions like anger or anxiety. The research in the field indicates that facial expressions, such as frowning or glaring, are commonly associated with anger, while anxious individuals may exhibit tense or fearful facial expressions [51].

In contrast, vocal signs like flat tone or a tense, rushed way of speaking can often reveal traces of depression or anxiety. The studies in the area have shown that individuals with depression regularly manifest a lower pitch, a slower speech, and a reduced variability in their voice. Similarly, diseases like anxiety can also manifest in speech patterns; for example, when individuals display higher pitch and faster speech rates [52].

On the other hand, a person's language behavior in speech or writing can provide deeper insight into complex struggles, such as suicidal thoughts, psychosis, physical pain, or substance use. In the same reasoning, linguistic analyses have found types of patterns like the increased use of words associated with the emotions of "anger" and "sadness," among individuals with suicidal thoughts. In addition, language problems are also key features of schizophrenia spectrum disorders and may indicate a higher risk of psychosis. Having that in mind, the following table (cf. Table 11) establishes a bridge between how each modality contributes to identifying different types of mental health indicators [53], [54].

Table 11 – DSM-5 Symptom Domains and Corresponding Multimodal Digital Phenotypes

DSM-5 Domain	Data Type Used	System Component
Depression	Text + Voice + Facial expression	Sentiment Analysis + Vocal Tone
Anxiety	Facial Expression + Voice	Microexpressions + Prosody
Anger	Facial Expression	Facial Expression Detection
Mania	Text + Facial Expression	Accelerated Language + Hyperexpressivity
Suicidal Ideation	Text	Semantic Analysis (LLM)
Psychosis	Text + Voice	Linguistic Patterns + Discourse Inconsistencies
Sleep Disorders	Text + Voice	Textual Statements + Vocal Fatigue
Somatic Symptoms	Text + Voice	Frequent Mentions of Pain/Discomfort
Repetitive Behaviors	Text	Lexical Repetition + Obsessive Phrases
Dissociation	Voice + Text	Vague Language + Monotonous Speech
Memory	Text + Facial Expression	Incomplete or Disorganized Sentences
Personality Functioning	Text	Discourse Coherence + Relational Content
Substance Use	Text	Direct/Indirect Mentions of Consumption

4.2.3 Study Participants

The validation phase involved a sample of adult participants, which included simulated sessions with volunteers devised to reproduce typical clinical-psychological interactions.

Table 12 – Inclusion Criteria

Criteria	Description
Age	≥18 years
Communication	Great verbal Communication
Consent	Pre-Agreement for session recording and survey filling

Table 13 – Exclusion Criteria

Criteria	Description
Cognitive capacity	Presence of cognitive disabilities that prevent completion of the survey
Data usage consent	Withdrawal of consent for using image, voice, and data for research purposes

In this stage, 20 volunteers were selected, allowing a statistically relevant analysis.

4.2.4 Experimental Procedure and Results

The experimental procedure follows a structured sequence of steps to ensure an efficient evaluation of the proposed system. Initially, prior to the session, each participant must complete a DSM-5 questionnaire, which documents symptoms, behaviors, and self-observed conduct in recent periods. Subsequently, the clinical session takes place.

The session is also aligned with the DSM-5 Symptom Measure. All the questions from the questionnaire are meant to be addressed during the appointment, and its data will be processed and collected by the three emotion recognition models, to be later utilized for results comparison and system validation.

Finally, upon the session conclusion, the report will be generated, and the results of the respective clinical domains will be available for analysis and study. These data will be compared against the alignment with the participant questionnaire, enabling the validation of the proposed methodology.

Following the described validation procedure, simulated sessions were carried out with the participation of 20 volunteers. This subsection presents the obtained results of the cross-validation between the developed system and the DSM-5 questionary responses, highlighting the evaluation performance metrics and the overall statistical results correlation between both.

For each of the 13 clinical domains defined by the DSM-5 scale, a binary classification ("Yes" or "No") was adopted based on the significant presence of symptoms, following the criteria described in sub-section 5.1 –responses with values equal to or greater than 2 were labeled as "Yes," while all others were labeled as "No” (cf. Table 14).

Table 14 – Cross-Validation Sample of System Predictions Against DSM-5 Responses

Volunteer	Domain	DSM Response (0–4)	Above the threshold? (YES/NO)	System Detected?	TP	FP	FN	TN
1	Depression	4	NO	YES	0	1	0	0
1	Anxiety	0	NO	NO	0	0	0	1
1	Anger	4	YES	YES	1	0	0	0
..	…	…	…	…	…	..	..	..
20	Memory	4	YES	YES	1	0	0	0
20	Personality Functioning	2	YES	YES	1	0	0	0
20	Substance Use	1	NO	YES	0	1	0	0

From the individual classification outcomes presented in the previous table (cf. Table 14), performance metrics were calculated per clinical domain, including accuracy, precision, recall, and F1-score. The Table 15 summarizes these metrics across the 13 DSM-5 domains, based on the data of the 20 study volunteers. The presented values were obtained from the cumulative counts of TP, FP, FN and TN attained during the validation process.

Table 15 – Performance Metrics by Clinical Domain (n = 20 participants)

Domain	TP	FP	FN	TN	Accuracy	Precision	Recall	f1-score
Depression	7	3	4	6	0,65	0,54	0,64	0,59
Anxiety	8	2	3	7	0,75	0,53	0,73	0,61
Anger	13	2	1	4	0,85	0,76	0,93	0,84
Mania	10	0	3	7	0,85	0,59	0,77	0,67
Suicidal Ideation	8	4	5	3	0,55	0,73	0,62	0,67
Psychosis	10	3	4	3	0,65	0,77	0,71	0,74
Sleep Disorders	7	0	4	9	0,8	0,44	0,64	0,52
Somatic Symptoms	10	1	2	7	0,85	0,59	0,83	0,69
Repetitive Behaviors	6	3	5	6	0,6	0,5	0,55	0,52
Dissociation	5	2	5	8	0,65	0,38	0,5	0,43
Memory	12	1	1	6	0,9	0,67	0,92	0,78
Personality Functioning	13	3	2	2	0,75	0,87	0,87	0,87
Substance Use	7	7	3	3	0,5	0,7	0,7	0,7

Clinical domains such as Memory, Anger, and Personality Functioning showed particularly high F1-scores, indicating that the multimodal input processed by the system were especially effective in capturing patterns associated with these symptoms. Conversely, domains like

Dissociation and Repetitive Behaviors showed lower overall performance, which may be attributed to the subtlety or variability of how these symptoms manifest across visual, vocal, and verbal cues (cf. Figure 17).

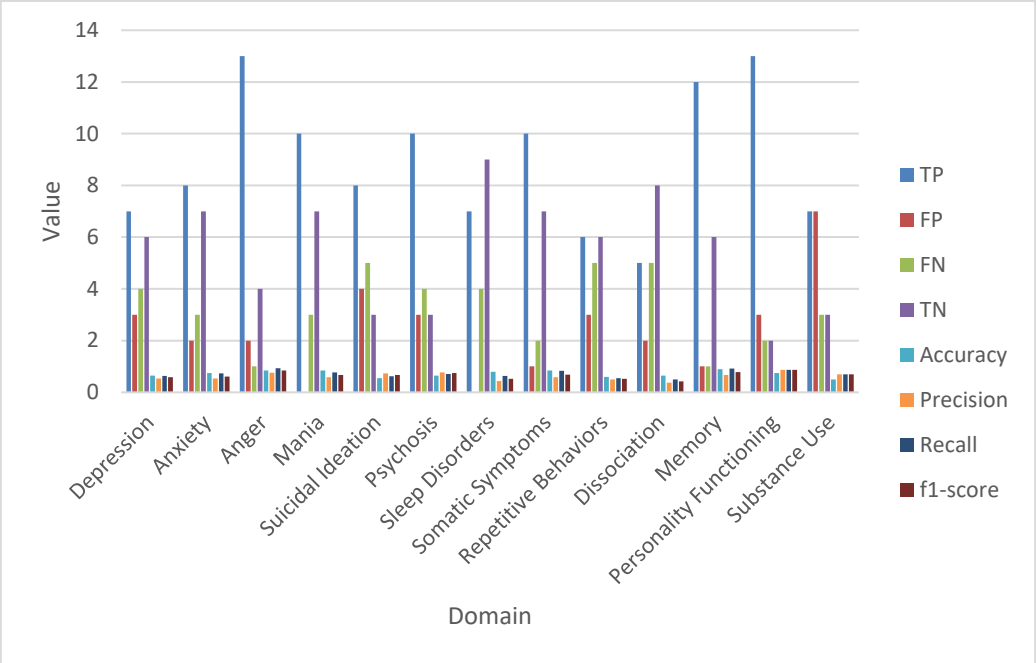

Figure 17 – Performance Metrics by Clinical Domain

Beyond the domain-level analysis, a global confusion matrix was constructed (cf. Table 16) by aggregating the values from all domains and participants. This matrix serves as the basis for assessing the overall relation between the developed system and the DSM-5 clinical tool.

Table 16 – Global Confusion Matrix

	DSM-5: YES	**DSM-5: NO**	**Total**
System: YES	116	31	147
System: NO	42	71	113
Total	158	102	260

The data established on the matrix was used to compute and evaluate the agreement between the system's predictions and the DSM-5 responses, while correcting for chance agreement. The coefficient is calculated based on the observed agreement – Po = 0.719 – and the expected agreement by chance – Pe = 0.514 –, resulting in a final Cohen's Kappa of 0.422.

According to Landis and Koch's scale [55] this value indicates a moderate level of agreement. This supports the system's potential to reliably approximate the preliminary diagnostic indications defined by the DSM-5, particularly considering the inherent complexity and subjectivity of mental health symptom detection through multimodal signals.

4.3 Discussion

The findings from the system experimental validation show a meaningful capacity to clinically identify relevant symptoms through the analysis of multimodal signals – facial expressions, vocal features, and textual content his is consistent with the system's objective for mental health symptom detection, where the fusion of multiple and different data inputs aims to bring a more comprehensive analysis of the complexity of human mental well-being.

The use of traditional evaluation metrics – accuracy, precision, recall, and the F1-score – granted a detailed assessment of the system's performance regarding the different DSM-5 domains. Globally, the results were promising. Domains like Memory, Anger, Personality Functioning, and Mania were particularly notable, showing consistently high F1-scores, exceeding in different cases the score of 0.75. These results align with previous studies that suggest that these types of symptoms are often associated with easier pattern detection in voice features, facial behavior, and language.

In contrast, performance levels in areas such as Repetitive Behaviors and Dissociation revealed more modest values. This might indicate that the current approach has limitations towards more subtle or less manifest symptoms. This aligns with existing research that highlights the difficulty in identifying disruptive mental conditions through external signals, pointing instead to the need for a more detailed clinical investigation.

The global confusion matrix revealed a good proportion of correct classifications, with 116 true positives and 71 true negatives. Cohen's Kappa coefficient, calculated at 0.42, indicates a modest level of agreement between the system and the clinical reference (DSM-5). Considering the subjective nature of many of the evaluated symptoms, the result was noteworthy and fairly consistent with other studies in the field, where Kappa values between 0.40 and 0.60 are considered acceptable indicators of preliminary clinical reliability.

Overall, the findings suggest that a system can have genuine potential to act as a clinical decision support tool, especially around initial screening, ongoing monitoring, or as a complement to clinical reasoning.

5 Conclusions

In this chapter, an overview of the study and the developed system is presented, followed by the objectives achieved. Subsequently, the study results, the limitations encountered, and suggestions for future work are offered.

5.1 Summary and Objectives Achieved

This study proposed the creation of an AI-powered system for emotion recognition to aid in mental health diagnosis and therapeutic support, following a multimodal approach incorporating facial expression recognition, speech emotion detection, and sentiment analysis on text data. The main goal was to equip mental health specialists with actionable, real-time information and comprehensive reports alongside conventional diagnosis techniques and to aid in therapeutic decision-making, especially in remote consultations. The central research question that drove this work was:

"How can an AI-powered system leverage emotion recognition techniques to improve mental health diagnostics and therapeutic interventions?"

The initial part of the study introduced theoretical and contextual aspects of emotion recognition systems which described main methodologies, evaluation metrics, challenges and their applications in mental health. A systematic review was performed to evaluate the relative performance between multimodal and unimodal approaches and their suitability for real-time clinical practice implementation. The review results demonstrated that multimodal systems have advantages in terms of detection robustness and accuracy through the combination of facial and vocal and textual modalities.

The following section described the design, development methods and materials of the proposed system using Design Science Research methodology. The selection of the datasets proved to be essential for achieving both model accuracy and generalization. The FER2013 dataset served as the facial emotion recognition dataset, containing standardized grayscale

48×48-pixel images that cover a wide range of emotions. The speech emotion recognition component received training from a dataset that combined CREMA-D, RAVDESS, SAVEE and TESS datasets which provided diverse speaker populations and emotional expressions and recording environments. The text sentiment analysis module required a dataset of anonymized psychologist – patient conversations to extract realistic linguistic patterns that match mental health contexts. The preprocessing phase for each dataset included cleaning and normalization followed by feature extraction and transformation into training-friendly formats.

The chapter that follows (Chapter 4) is directly related to the achievement of the objectives defined in Chapter 1 (Section 1.4). The objectives defined to answer the proposed research question led to the design and implementation of the system, and their fulfillment can be validated as follows:

1. Multimodal Emotion Recognition System – Achieved through the integration of three emotion analysis modules (facial expression, speech, and text sentiment analysis models).

2. Professional Diagnostic Reports – Implemented in the back-end, responsible for generating the necessary data and later feeding it to the LLM, which is accountable for generating the report; and front-end architecture, responsible for displaying the generated session summaries and interactive insights.

3. Personalized Therapeutic Suggestions – Addressed via the generated reports and the interactive chatbot that enables professionals to explore and search for patterns more effectively. Furthermore, the LLM was fine-tuned with professional reports and domain-specific resources, which improved its capacity to give contextually relevant therapeutic insights in the field of mental health.

4. Real-Time Notifications – Delivered by the real-time workflow, which monitors the emotional changes during the consultations.

5. Integration with Remote Video Call Platforms – While a standalone web application has been implemented and validated, the plug-in integration for platforms such as Zoom and Microsoft Teams remains as future work .

The system was validated through an experimental procedure that compared its predictions against the DSM-5 Self-Rated Level 1 Cross-Cutting Symptom Measure, used as a clinical reference standard. The validation process involved twenty adult participants who underwent simulated clinical scenarios, with each session beginning with completion of the DSM-5 questionnaire. During these sessions, all questions from the scale were addressed, and the corresponding audio, video, and text data were processed in real time by the system. The results of the questionnaire – using binary classification to indicate the significant presence or absence of symptoms –, were compared with the system outputs for each of the thirteen DSM-5 domains.

The results of this validation showed a moderate level of agreement with the DSM-5, with Cohen's Kappa coefficient at 0.42. This value matches findings from other studies in the field. Some areas, like Memory, Anger, and Personality Functioning, reached high F1-scores of 0.78, 0.84, and 0.87, respectively. This confirms that the system effectively captures these patterns through multimodal analysis. In contrast, areas like Dissociation and Repetitive Behaviors produced lower scores. This suggests that detecting symptoms that appear subtly in external signals is more challenging. Overall, the system recorded 116 true positives and 71 true negatives in the aggregated confusion matrix. This reflects its potential for preliminary screening and supporting clinical decisions.

From the beginning, the study aimed to achieve five specific goals. The development of a multimodal emotion recognition system, the generation of professional clinical reports, the integration of a personalized therapeutic suggestion module, and the implementation of a real-time notification system were all successfully completed. However, integrating the solution as a plugin for platforms like Zoom or Microsoft Teams – which intended to enable the use of the tool during telemedicine sessions – was not accomplished in this phase and remains a task for future work.

5.2 Limitations and Future Work

During the development and validation of the proposed multimodal emotion recognition system, several limitations were identified that should be considered for future improvements.

Firstly, the experimentation and evaluation phases were conducted in a controlled and simulated environment, which does not fully replicate the complexity and variability of real-world clinical settings. Factors such as unstable internet connections, background noise, varied lighting conditions, and spontaneous patient behaviors may impact performance. Real-world trials with mental health professionals and patients will be essential to validate the system's robustness, usability, and overall impact in practical scenarios.

Secondly, the system's validation focused exclusively on quantitative performance metrics obtained from benchmark datasets. While these metrics are essential to assess the technical accuracy of the models, they do not capture aspects such as usability, interpretability, and actual effectiveness in supporting clinical decision-making. Without direct feedback from mental health professionals and real patient interactions, it is not possible to fully assess the system's practical value in therapeutic contexts. Future studies should include user-centered evaluations and clinical pilot tests to address this gap.

Additionally, although the back-end architecture was designed with extensibility in mind and already includes the necessary foundations to integrate real-time plugins for Microsoft Teams or Zoom, these integrations were not implemented in the current version. However, it is important to note that both Microsoft Teams and Zoom impose strict security, compliance, and data privacy policies for third-party integrations. These platforms require certification

processes, adherence to GDPR standards, and rigorous approval workflows before allowing external plugins to interact with live video or audio streams. Such obligations introduced delays, additional development effort, and even blocking points, depending on how sensitive data is processed. As a result, the project delivery timeline was not compatible with the integration approval timeline, which justifies why this step is proposed as future work.

Addressing these limitations and implementing the proposed improvements will enhance the system's reliability, fairness, and clinical relevance, paving the way for broader adoption in mental health diagnostics and therapeutic interventions.

References

[1] A. Chakraborty *et al.*, "A Framework of Intelligent Mental Health Monitoring in Smart Cities and Societies," *IETE J Res*, vol. 70, no. 2, pp. 1328–1341, Feb. 2024, doi: 10.1080/03772063.2023.2171918.

[2] Y. Liang, X. Zheng, and D. D. Zeng, "A survey on big data-driven digital phenotyping of mental health," *Information Fusion*, vol. 52, pp. 290–307, Dec. 2019, doi: 10.1016/J.INFFUS.2019.04.001.

[3] K. Pandey *et al.*, "Mental Health Issues During and After COVID-19 Vaccine Era," *Brain Res Bull*, vol. 176, pp. 161–173, Nov. 2021, doi: 10.1016/J.BRAINRESBULL.2021.08.012.

[4] C. Wu, D. Davaasuren, T. Shafir, R. Tsachor, and J. Z. Wang, "Bodily expressed emotion understanding through integrating Laban movement analysis," *Patterns*, vol. 4, no. 10, p. 100816, Oct. 2023, doi: 10.1016/J.PATTER.2023.100816.

[5] S. Zhou, X. Wu, F. Jiang, Q. Huang , and C. Huang, "Emotion Recognition from Large-Scale Video Clips with Cross-Attention and Hybrid Feature Weighting Neural Networks," *International Journal of Environmental Research and Public Health 2023, Vol. 20, Page 1400*, vol. 20, no. 2, p. 1400, Jan. 2023, doi: 10.3390/IJERPH20021400.

[6] S. M. Head TA, C. Manning, and J. Cho, "CS 224N: Natural Language Processing with Deep Learning," Stanford / Spring 2024. Accessed: Nov. 02, 2024. [Online]. Available: https://web.stanford.edu/class/cs224n/

[7] Y.-G. Wu, Z.-E. Jin, and C.-L. Lin, "Emotion Detection by Speech and Voice Tone," in *Proceedings of the 25th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2021)*, Tainan City, Jul. 2021. Accessed: Apr. 22, 2025. [Online]. Available: https://www.researchgate.net/profile/Yung-Gi-Wu/publication/358736676_Emotion_Detection_by_Speech_and_Voice_Tone/links/6212594ff02286737cafe78f/Emotion-Detection-by-Speech-and-Voice-Tone.pdf

[8] J. Venable, J. Pries-Heje, and R. Baskerville, "FEDS: a Framework for Evaluation in Design Science Research," *European Journal of Information Systems*, vol. 25, no. 1, pp. 77–89, Jan. 2016, doi: 10.1057/EJIS.2014.36.

[9] J. R. Venable, J. Pries-Heje, and R. L. Baskerville, "Choosing a Design Science Research Methodology," *ACIS 2017 Proceedings*, Jan. 2017, Accessed: Jan. 03, 2025. [Online]. Available: https://aisel.aisnet.org/acis2017/112

[10] C. Sohrabi *et al.*, "PRISMA 2020 statement: What's new and the importance of reporting guidelines," *International Journal of Surgery*, vol. 88, p. 105918, Apr. 2021, doi: 10.1016/J.IJSU.2021.105918.

[11] A. Van Der Merwe, “Guidelines for Conducting Design Science Research in Information Systems,” *SACLA*, pp. 1–17, 2019, Accessed: Feb. 09, 2025. [Online]. Available: http://sacla2019.unisa.ac.za/Preprints/FP_van%20der%20Merwe_Design%20Science%20Research.pdf

[12] C. Zhu, T. Ding, and X. Min, “Emotion Recognition of College Students Based on Audio and Video Image,” *Traitement du Signal*, vol. 39, no. 5, pp. 1475–1481, Oct. 2022, doi: 10.18280/TS.390503.

[13] K. Dheeraj and T. Ramakrishnudu, “Negative emotions detection on online mental-health related patients texts using the deep learning with MHA-BCNN model,” *Expert Syst Appl*, vol. 182, p. 115265, Nov. 2021, doi: 10.1016/J.ESWA.2021.115265.

[14] H. Xu, X. Wu, and X. Liu, “A measurement method for mental health based on dynamic multimodal feature recognition,” *Front Public Health*, vol. 10, Dec. 2022, doi: 10.3389/FPUBH.2022.990235.

[15] X. Peng, “Research on emotion recognition based on deep learning for mental health,” *Informatica (Slovenia)*, vol. 45, no. 1, pp. 127–132, 2021, doi: 10.31449/INF.V45I1.3424.

[16] P. Meshram and R. K. Rambola, “Diagnosis of depression level using multimodal approaches using deep learning techniques with multiple selective features.,” *Expert Syst*, vol. 40, no. 4, pp. 1–13, May 2023, doi: 10.1111/EXSY.12933.

[17] C. Zhu, T. Ding, and X. Min, “Emotion Recognition of College Students Based on Audio and Video Image,” *Traitement du Signal*, vol. 39, no. 5, pp. 1475–1481, Oct. 2022, doi: 10.18280/TS.390503.

[18] C. Li, M. Yang, Y. Zhang, and K. W. Lai, “An Intelligent Mental Health Identification Method for College Students: A Mixed-Method Study,” *Int J Environ Res Public Health*, vol. 19, no. 22, Nov. 2022, doi: 10.3390/IJERPH192214976.

[19] A. Davodabadi, B. Daneshian, S. Saati, and S. Razavyan, “Mathematical model and artificial intelligence for diagnosis of Alzheimer’s disease,” *Eur Phys J Plus*, vol. 138, no. 5, May 2023, doi: 10.1140/EPJP/S13360-023-04128-5.

[20] S. Zhou, X. Wu, F. Jiang, Q. Huang , and C. Huang, “Emotion Recognition from Large-Scale Video Clips with Cross-Attention and Hybrid Feature Weighting Neural Networks,” *Int J Environ Res Public Health*, vol. 20, no. 2, Jan. 2023, doi: 10.3390/IJERPH20021400.

[21] N. Shanthi *et al.*, “An integrated approach for mental health assessment using emotion analysis and scales,” *Healthc Technol Lett*, 2022, doi: 10.1049/HTL2.12040.

[22] Z. Fei *et al.*, “Deep convolution network based emotion analysis towards mental health care.,” *Neurocomputing*, vol. 388, pp. 212–227, May 2020, doi: 10.1016/J.NEUCOM.2020.01.034.

[23] J. Aina, O. Akinniyi, M. M. Rahman, V. Odero-Marah, and F. Khalifa, “A Hybrid Learning-Architecture for Mental Disorder Detection Using Emotion Recognition,” *IEEE Access*, vol. 12, pp. 91410–91425, 2024, doi: 10.1109/ACCESS.2024.3421376.

[24] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, and M. Al-Hammadi, “A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition,” *Information Fusion*, vol. 53, pp. 209–221, Jan. 2020, doi: 10.1016/J.INFFUS.2019.06.019.

[25] C. Wu, D. Davaasuren, T. Shafir, R. Tsachor, and J. Z. Wang, “Bodily expressed emotion understanding through integrating Laban movement analysis,” *Patterns (N Y)*, vol. 4, no. 10, Oct. 2023, doi: 10.1016/J.PATTER.2023.100816.

[26] S. Hossain, S. Umer, R. K. Rout, and H. Al Marzouqi, “A Deep Quantum Convolutional Neural Network Based Facial Expression Recognition For Mental Health Analysis,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 1556–1565, 2024, doi: 10.1109/TNSRE.2024.3385336.

[27] X. Zhang, B. Li, and G. Qi, “A multi-featured expression recognition model incorporating attention mechanism and object detection structure for psychological problem diagnosis.,” *Physiol Behav*, vol. 280, p. N.PAG-N.PAG, Jun. 2024, doi: 10.1016/J.PHYSBEH.2024.114561.

[28] Y. Zhao, Z. Liang, J. Du, L. Zhang, C. Liu, and L. Zhao, “Multi-Head Attention-Based Long Short-Term Memory for Depression Detection From Speech,” *Front Neurorobot*, vol. 15, Aug. 2021, doi: 10.3389/FNBOT.2021.684037.

[29] C. Doğdu, T. Kessler, D. Schneider, M. Shadaydeh, and S. R. Schweinberger, “A Comparison of Machine Learning Algorithms and Feature Sets for Automatic Vocal Emotion Recognition in Speech.,” *Sensors (14248220)*, vol. 22, no. 19, pp. 7561–7561, Oct. 2022, doi: 10.3390/S22197561.

[30] Y. Alemu, H. Chen, C. Duan, D. Caulley, R. I. Arriaga, and E. Sezgin, “Detecting Clinically Relevant Emotional Distress and Functional Impairment in Children and Adolescents: Protocol for an Automated Speech Analysis Algorithm Development Study,” *JMIR Res Protoc*, vol. 12, 2023, doi: 10.2196/46970.

[31] R. C. Aguilera, A. A. Banda, M. P. Ortiz, and J. P. Ortiz, “DYNAMIC HYPERLEDGER NFT ON FEDERATED LEARNING FOR PSYCHIATRIC SERVICES IN THE COVID-19 TIMES.,” *Fractals*, vol. 31, no. 7, pp. 1–14, 2023, doi: 10.1142/S0218348X23500718.

[32] K. Dheeraj and T. Ramakrishnudu, “Negative emotions detection on online mental-health related patients texts using the deep learning with MHA-BCNN model.,” *Expert Syst Appl*, vol. 182, p. N.PAG-N.PAG, Nov. 2021, doi: 10.1016/J.ESWA.2021.115265.

[33] A. B. S. Rahman, H. T. Ta, L. Najjar, A. Azadmanesh, and A. S. Gönul, “DepressionEmo: A novel dataset for multilabel classification of depression emotions.,” *J Affect Disord*, vol. 366, pp. 445–458, Dec. 2024, doi: 10.1016/J.JAD.2024.08.013.

[34] D. Kodati and R. Tene, “Advancing mental health detection in texts via multi-task learning with soft-parameter sharing transformers.,” *Neural Comput Appl*, pp. 1–34, 2024, doi: 10.1007/S00521-024-10753-7.

[35] C. Li, Y. Mao, Q. Huang, W. Xie, X. He, and J. Wu, “A Real-Time Emotion-Aware System Based on Wireless Body Area Network for IoMT Applications,” *IEEE Internet Things J*, 2024, doi: 10.1109/JIOT.2024.3458976.

[36] F. Gioia, A. Greco, A. L. Callara, and E. P. Scilingo, “Towards a Contactless Stress Classification Using Thermal Imaging,” *Sensors 2022, Vol. 22, Page 976*, vol. 22, no. 3, p. 976, Jan. 2022, doi: 10.3390/S22030976.

[37] A. Merla and G. L. Romani, “Thermal signatures of emotional arousal: a functional infrared imaging study,” *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2007, pp. 247–249, 2007, doi: 10.1109/IEMBS.2007.4352270.

[38] C. L. Lee, W. Pei, Y. C. Lin, A. Granmo, and K. H. Liu, “Emotion Detection Based on Pupil Variation,” *Healthcare (Basel)*, vol. 11, no. 3, Feb. 2023, doi: 10.3390/HEALTHCARE11030322.

[39] A. El Abbaoui, D. Sodoyer, and F. Elbahhar, “Contactless Heart and Respiration Rates Estimation and Classification of Driver Physiological States Using CW Radar and Temporal Neural Networks,” *Sensors (Basel)*, vol. 23, no. 23, Dec. 2023, doi: 10.3390/S23239457.

[40] P. R. Jain, S. M. K. Quadri, and A. Khattar, “EM-UDA: Emotion Detection using Unsupervised Domain Adaptation for Classification of Facial Images,” *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3467990.

[41] K. Denecke and E. Gabarron, “The ethical aspects of integrating sentiment and emotion analysis in chatbots for depression intervention,” *Front Psychiatry*, vol. 15, 2024, doi: 10.3389/FPSYT.2024.1462083.

[42] J. Singh, L. B. Saheer, and O. Faust, “Speech Emotion Recognition Using Attention Model,” *Int J Environ Res Public Health*, vol. 20, no. 6, Mar. 2023, doi: 10.3390/IJERPH20065140.

[43] J. Oh *et al.*, “Development of depression detection algorithm using text scripts of routine psychiatric interview,” *Front Psychiatry*, vol. 14, 2024, doi: 10.3389/FPSYT.2023.1256571.

[44] A. Pollock and E. Berge, “How to do a systematic review,” *International Journal of Stroke*, vol. 13, no. 2, pp. 138–156, Feb. 2018, doi:

10.1177/1747493017743796/ASSET/IMAGES/LARGE/10.1177_1747493017743796-FIG6.JPEG.

[45] A. Adeleye, S. Madanian, and O. Adeleye, “Emotion Variation Detection in Discrete English Speech: A Wavelet Transform Use Case in Mental Health Monitoring,” *ACM International Conference Proceeding Series*, pp. 115–119, Jan. 2024, doi: 10.1145/3641142.3641167.

[46] D. Caulley *et al.*, “Objectively Quantifying Pediatric Psychiatric Severity Using Artificial Intelligence, Voice Recognition Technology, and Universal Emotions: Pilot Study for Artificial Intelligence-Enabled Innovation to Address Youth Mental Health Crisis,” *JMIR Res Protoc*, vol. 12, no. 1, 2023, doi: 10.2196/51912.

[47] Y. W. Leung *et al.*, “Therapist Feedback and Implications on Adoption of an Artificial Intelligence-Based Co-Facilitator for Online Cancer Support Groups: Mixed Methods Single-Arm Usability Study,” *JMIR Cancer*, vol. 9, 2023, doi: 10.2196/40113.

[48] European Parliament and Council of the European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation, GDPR)*. European Union, 2016. Accessed: May 06, 2025. [Online]. Available: https://gdpr-info.eu/

[49] European Commission, *The AI Act Explorer | EU Artificial Intelligence Act*. 2024. Accessed: May 06, 2025. [Online]. Available: https://artificialintelligenceact.eu/ai-act-explorer/

[50] P. Kruchten, “Architectural Blueprints-The ‘4+1’ View Model of Software Architecture,” *IEEE Softw*, vol. 12, no. 6, pp. 42–50, 1995.

[51] K. Cherry, “The 6 Types of Basic Emotions,” verywellmind. Accessed: Aug. 18, 2025. [Online]. Available: https://www.verywellmind.com/an-overview-of-the-types-of-emotions-4163976?utm_source=chatgpt.com

[52] D. Shin *et al.*, “Detection of Minor and Major Depression through Voice as a Biomarker Using Machine Learning,” *J Clin Med*, vol. 10, no. 14, p. 3046, Jul. 2021, doi: 10.3390/JCM10143046.

[53] L. Baklund, J. I. Røssberg, and P. Møller, “Linguistic markers and basic self-disturbances among adolescents at risk of psychosis. A qualitative study,” *EClinicalMedicine*, vol. 55, Jan. 2023, doi: 10.1016/j.eclinm.2022.101733.

[54] M. F. Dobbs *et al.*, “Linguistic correlates of suicidal ideation in youth at clinical high-risk for psychosis,” *Schizophr Res*, vol. 259, p. 20, Sep. 2023, doi: 10.1016/J.SCHRES.2023.03.014.

[55] C. Díaz, G. González, L. I. Jara-Ramírez, and J. A. Muñoz-Parra, “Validation of a classroom management questionnaire for pre and inservice teachers of English,” *Revista*

Colombiana de Educacion, no. 75, pp. 263–285, Jul. 2018, doi: 10.17227/RCE.NUM75-8111.