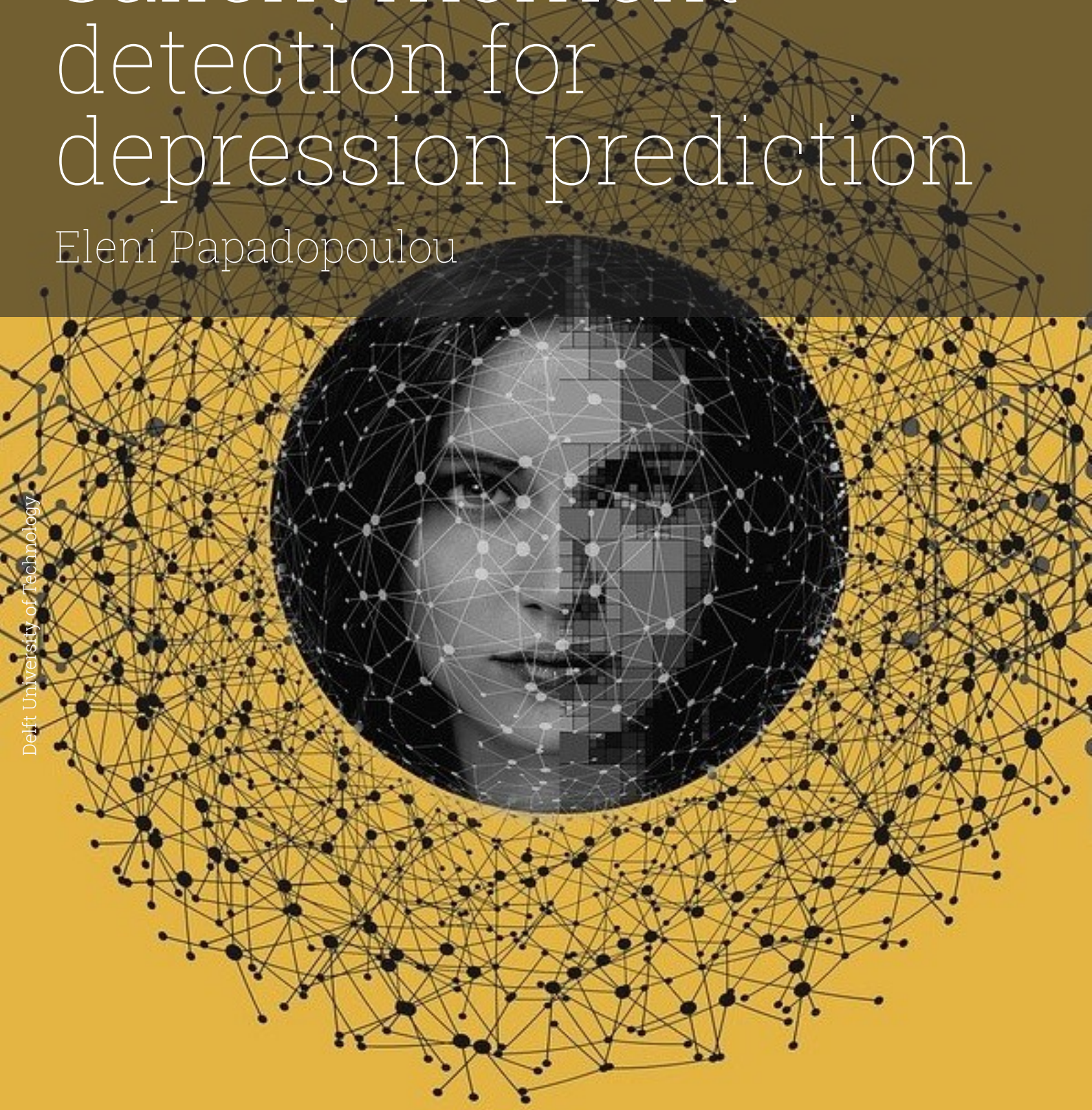
Salient moment detection for depression prediction

Eleni Papadopoulou



Delft University of Technology

TU Delft

Salient moment detection for depression prediction

by

Eleni Papadopoulou

Eleni
Papadopoulou

Student number: 5848148
Project Duration: 04, 2024 - 05, 2025
Faculty: Electrical Engineering, Mathematics and Computer Science, Delft
Supervisors: C. R. M. M. Oertel,
C. A. Raman,
A. Axelsson.

TU Delft

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Catharine Oertel, Chirag Raman and Agnes Axelsson. Their patient guidance, insightful feedback, and support throughout this last year helped me grow and finish this project. Also, I would like to thank all the participants of my experiments, because without them, this work would not have been possible. Also, I would like to thank my friends for all the times that they offered me laughter when I needed it the most, for celebrating my work and for all our conversations about this project and the interest they showed.

A special thank you to my amazing parents for their financial and emotional support throughout, not just this thesis but my whole life. To my sibling Aggeliki that is always here for me and makes me laugh, gives me advice and offers me unlimited emotional support. Last but not least, to my lovely boyfriend Najib for the endless talks, laughs, support and insightful advice for this project. Your love helped me grow and believe in myself.

Σας ευχαριστώ

Eleni Papadopoulou
Delft, May 2025

Abstract

Early detection of depression is crucial in mental healthcare. Augmenting depression diagnosing with AI seems to be promising in detecting depression from subtle non-verbal cues and early signs that can be missed from domain experts. For this to be achieved, AI procedures and decision processes need to be interpretable to humans. In this thesis, we use and evaluate a saliency-based explainability framework for a multi-modal depression-prediction model and validate its outputs through human judgment. The multimodal input is created by combining high level facial features extracted from Action-Unit via a 1D-CNN and high level vocal features extracted from log-mel spectrograms via a modified AlexNet. Then a simple Feed Forward Network is used as a classification predictor for 3.5 second segments.

To assess whether these AI-flagged moments align with human reasoning, 17 lay participants viewed thirty 8.5-second clips (half depressed, half non-depressed). For each clip they (1) rated depression confidence on a 1–10 scale, (2) selected the single frame they found most influential, and (3) described the facial or vocal cues that informed their choice. The goal is for the participants to give us an insight on what the model may be 'seeing'. So we ask them to tell us what facial and voice features they observed in their influential moments. From those experiments, we gained some useful insights to the model. The results show that participants observations in non-verbal cues are valuable and align with literature findings. And we find that there is alignment in participant's observations on their own influential moment and on the model's salient moment when the salient moment is correctly classified by the model and they do not align when the salient moment is wrongly classified by the model. These findings suggest that humans and the model value similar cues to make the correctly classify depression, and help to enhance the interpretability of AI models.

Contents

1

Introduction

Depressive disorder (also known as depression) is one of the prevalent mental health disorders. At the global level, more than 300 million people are estimated to suffer from depression, equivalent to 4.4% of the world's population World Health Organization 2017. The main characteristic of depression is persistently low mood or a lack of interest and enjoyment for an extended period. Depression is not associated with typical fluctuations in mood or feeling, as its severity can affect various aspects of the individual's life such as relationships or professional settings. The number of people diagnosed with depression or experiencing depressive symptoms has increased steadily in the last century Wilson and Dumornay 2022. Although depression is a leading cause of global disease burden, many individuals do not receive adequate treatment, resulting in higher risks of chronicity and relapse Thornicroft et al. 2017, Mekonen et al. 2022. Moreover, diagnosis in mental health remains challenging . Traditional frameworks, such as the *Diagnostic and Statistical Manual of Mental Disorders*, rely on identifying a minimum number of core symptoms over a specified period. Furthermore, misdiagnosis rates for certain mental health conditions (e.g., bipolar disorder) can be as high as 55–76% Byeon 2023, underscoring the limitations of current diagnostic methods and the need for innovative solutions.

In recent years, the use of new technological advancements for medical purposes has become increasingly common (Junaid et al., 2022). Early detection of depression is crucial because it allows patients to receive treatment sooner. One way to achieve this is by recognizing subtle signs that human diagnosticians might miss. Research indicates that subtle non-verbal cues—such as minor changes in facial expressions, tone of voice, and body language—are often difficult for clinicians to observe during routine assessments. For example, Cummins et al. 2015 discuss how nuanced variations in vocal prosody and speech patterns, which can serve as early markers of depression, are frequently overlooked in clinical practice. That is because these subtle cues can be brief and can vary greatly between individuals, making it difficult to consistently identify early indicators of depression during those routine assessments. Early identification of these cues is crucial, as alerting both the therapist and the patient can prevent relapse and improve long-term outcomes. In response to these challenges, researchers are exploring innovative strategies—such as patient self-monitoring tools (Onnela and Rauch, 2016) and hybrid systems (Balcombe and De Leo, 2021) that integrate clinician insights with AI analysis—to better capture the diverse ways depression manifests. At the same time, AI-based systems that analyze text, audio, and video data have been proven to be powerful tools for diagnosing and predicting depression, showing the potential to surpass traditional methods in the future, in terms of accuracy and efficiency (Shatte, Hutchinson, and Teague, 2019). However, many of these models, especially deep neural networks, operate as black boxes, offering high accuracy at the cost of interpretability (Shah and Konda, 2021). In the context of clinical decision making, where trust and transparency are paramount, this lack of clarity can hinder adoption and raise ethical concerns.

Explainable Artificial Intelligence (XAI) aims to address these challenges by making complex models more transparent without substantially compromising their predictive performance. Explainability is especially valuable in mental health: not only does it help clinicians understand why a model considers someone likely to be depressed, but it also potentially fosters trust among patients, who are more

likely to accept AI-driven insights if they know how those insights were arrived at (Woodcock et al., 2021). In mental health diagnostics—particularly for depression— it is well established that cues, such as shifts in vocal intonation, facial expressions, and body language, are inherently dynamic and are key indicators that evolve over time (Mundt, Vogel, et al., 2012; L.-S. A. Low et al., 2010). However, many existing XAI methods in this domain are primarily static, failing to fully capture how these cues evolve over as can be observed by the comprehensive review by Guidotti et al. (2018). Most studies to date either omit explainability or use generic feature-attribution techniques that ignore sequence order, issues mentioned in the works of Arik and Pfister (2021) and Imans et al. (2024). Traditional XAI methods like LIME (Ribeiro, Singh, and Guestrin, 2016), SHAP (Lundberg and Lee, 2017) have been applied to depression prediction (Byeon, 2023), but they treat input features as an unordered set, thus losing the narrative of change over time. In a clinical context, this can be problematic – we want to know not just which behaviors indicate depression, but when and how those behaviors manifest during an interaction.

In this research, we aim to address this problem by utilizing a saliency-based explainability method proposed by Raman et al. (2024). To implement this approach, we model the patient's input as a time-series by using a sliding window technique to segment the data into manageable intervals. Then we calculate the saliency of each segment comparing the change of the entropy from the previous segment, something that quantifies how much that particular interval (from the one segment to the other) contributes to the overall model prediction. Saliency, in this context, refers to the degree of influence that each segment exerts on the model's output. By assigning a saliency score to every segment, we can determine which parts of the patient's input are most critical in driving the model's classification decisions. This approach leverages the temporal dynamics of audio-visual data by identifying critical moments over time, thereby capturing how important cues evolve and influence the model's decisions - an insight that static methods may overlook. This is helpful because by highlighting the specific intervals that are the most critical to the model's classification decisions, we have a more intuitive explanation that can directly inform clinical decision-making in mental health. In contrast, most XAI techniques previously used for depression prediction are post-hoc feature attribution methods or simple attention-based interpretations, which are largely static in nature (Guidotti et al., 2018). Attention mechanisms have also been used to weight different time frames, which is a form of temporal explanation. While they highlight time steps that are important, it does not provide more a more granular explanation. In contrast, our saliency-based approach not only leverages temporal changes in audiovisual data but also quantifies the contribution of each segment—by, for example, measuring changes in entropy between adjacent segments—to the overall prediction. This can provide a more granular and interpretable explanation of the model's decision-making process, potentially enabling clinicians and patients to pinpoint exactly when key depressive cues manifest during an interaction. However, similarly to the attention mechanism techniques, while our approach identifies the salient moments that influence the model's decision, it does not specify which modality or precise feature within those moments is responsible.

Consequently, to further explore these temporal explanations, we conduct experiments with lay participants, asking them to review the model's highlighted audio-visual segments and answer targeted questions regarding the presence of potential depression cues. Additionally, we examine not only the participants' ability to recognize signs of depression, but also the extent to which their perception of influential cues and moment's aligns with the model's. This aims to enhance the interpretation of AI systems and their internal decision processing. This human validation is especially important given that in many studies on automated depression detection, the focus has been primarily on improving classification accuracy, often at the expense of a thorough evaluation of the interpretability of AI-generated explanations. Most existing work presents AI-highlighted segments in isolation without rigorously comparing them against human-selected cues or expert annotations. However, this kind of one-to-one comparison is crucial to ensure the AI is focusing on legitimate indicators of depression rather than fake patterns. For example, while studies like those by K. Yang et al. 2023 and L. Zhang et al. 2025 have made initial strides by introducing human-annotated explanations, these efforts remain isolated and do not fully integrate the temporal dimension of non-verbal behavior. Consequently, it remains unclear whether the model's highlighted moments genuinely correspond to the nuanced, time-varying signs of depression recognized by clinicians or lay evaluators.

Motivated by these considerations, this thesis analyzes whether it is possible to enhance the interpretability of AI-based medical diagnosis for depression through the identification of salient moments

from video and audio data and subjecting them to human evaluation. Specifically, the central research question is: *"To what extent can interpretability in AI-based medical diagnosis for depression be improved by identifying the salient moments in video and audio data and investigating them through human evaluation?"*

By addressing this question, we aim to determine whether the model's highlighted segments correspond to meaningful indicators that humans can recognize, and whether this alignment enhances clinicians' and patients' understanding of the rationale behind a depression diagnosis. To investigate further, we established the following sub-questions:

1. To what extent can humans identify depression from short video clips?
2. To what extent do human-identified salient moments align with the model's salient moments?
3. What facial/voice features do humans identify in AI-selected salient moments and in their own

By addressing these subquestions, we aim to establish a comprehensive evaluation approach that not only evaluates the model's predictive capabilities but also deepens our understanding of its explainability.

2

Background

2.1. Depression Prediction

As noted above, depression increasingly affecting our society having a significant impact to public health (World Health Organization et al., 2017). For this reason, using artificial intelligence technology for depression detection has become more prominent in recent years (Karimian et al., 2025). Machine learning methods are frequently used to analyze text data from social media, electronic health records, and patient self-reports Shatte, Hutchinson, and Teague 2019. Studies have shown that social media posts, on platforms like Twitter and Facebook, can be effectively used to identify individuals' signs of depression, potentially offering a non-invasive and continuous monitoring tool Guntuku et al. (2019). Additionally, various data modalities have been used in AI models such as Functional Magnetic Resonance Imaging (fMRI) Mousavian et al. (2021), genetic and biomarker data Gu, Ming, and Xie 2023, mobile phone data (Digital Phenotyping) Ware et al. 2020 etc. Frequently used data modalities used for depression prediction are questionnaires, electroencephalograms (EEG), and video recordings. For self-report questionnaires, prediction approaches typically involve traditional machine learning algorithms—such as logistic regression, support vector machines, and random forests (Yasin et al., 2023). In EEG-based prediction the process usually begins with advanced signal processing to extract relevant features (Choubey and Pandey, 2019). These features are then fed into classification algorithms such as support vector machines, random forests, or deep neural networks—including convolutional and recurrent architectures—to capture the temporal and spectral characteristics associated with depression (Yasin et al., 2023). In the case of video recordings, predictive methods borrow computer vision techniques for scanning for behavior signals. In this category, methods often rely on deep models of learning such as convolutional neural networks (CNNs) for extracting spatial features and recurrent neural networks (RNNs) or spatiotemporal models for handling facial expressions, body postures, and micro-expressions' temporal dynamics. Such models are well equipped to pick out subtle non-verbal signals that may point to depressive symptoms (Joshi and Kanoongo, 2022). Our research narrows its focus to leveraging audiovisual data as input for depression prediction. Ground truth labels for depression are derived from the Patient Health Questionnaire (PHQ), a widely validated instrument that provides a reliable measure of depression severity.

2.1.1. Personal Health Questionnaire Depression Scale (PHQ)

AI models that use questionnaire data have demonstrated good results in predicting depression. Commonly used self-report tools for evaluating depression symptoms are the Patient Health Questionnaire (PHQ-9 or PHQ-8) (Kroenke, Strine, et al., 2009) and the Beck Depression Inventory (BDI) (Beck, 1996). The PHQ questionnaire has proved to be a reliable and valid measure of depression severity according to Kroenke, Spitzer, and Williams 2001 and is widely used in research on depression prediction. It is used as the sole input in models (Jin et al., 2015), but most often, it is used in a hybrid setting, combined with other data like clinical or sociodemographic data (Hornstein et al., 2021; L. Yang et al., 2017; Jordan, Shedden-Mora, and Löwe, 2018) etc. The PHQ-8 derives from the PHQ-9, excluding

the 9th question regarding self-harm or suicide, focusing only on the other 8 symptoms of depression. This is because studies have shown that patients with or without diagnosed psychiatric illness can have passive thoughts of death without being in immediate risk of suicide, leading to a lack of clarity regarding what Item 9 of PHQ-9 is assessing (Razykov et al., 2012). In the DAIC-WOZ dataset that has been used in this project, patients are labeled as depressed or non-depressed based on their scores on the PHQ-8 questionnaire. Participants are classified as depressed if their PHQ-8 score is 10 or above, and as non-depressed if it is below 10.

2.1.2. Depression prediction from Audiovisual data

Audiovisual data is also a crucial source of information for depression prediction tasks since it provides non-verbal cues such as facial expressions, eye movements, and body language. For example, Girard, Cohn, et al. (2014) found that visual cues—such as subtle changes in facial expression and reduced eye contact—can accurately predict depressive symptoms. Early multimodal approaches combined facial and vocal signals to improve detection (Scherer, Stratou, and Morency, 2013; Gupta et al., 2014). Recent advances have further improved predictive performance by leveraging deep learning on audiovisual data. One such study used facial action analysis together with vocal prosody to automatically distinguish depressed patients from non-depressed (Z. Jiang et al., 2020). Othmani, Zeghina, and Muzammel (2022) developed a deep neural network model fusing facial and speech features as input for predicting depression relapse from audiovisual cues. Moreover, multimodal approaches that combine audiovisual data with other modalities have been developed to enhance prediction accuracy, including text (L. Yang et al., 2017) and EEG (Song et al., 2022). Typically classification accuracy rates fall in the 70–80% range for distinguishing depressed vs. non-depressed subjects based on audiovisual data as mentioned in the work of Girard and Cohn (2015). This is a substantial level of accuracy given the complexity of depression's presentation, and it approaches the performance one might expect from screening questionnaires. Some studies report even higher performance with advanced techniques – for instance, a recent deep learning approach using attention-based feature fusion from Mahayossanunt et al. (2023) obtained nearly 91.67% accuracy on a depression dataset. Overall, the convergence of findings – from small-scale experiments to large multimodal challenges – provides strong evidence that audiovisual markers are valid indicators of depression that algorithms can learn to recognize.

Building on this foundation, our work utilizes audiovisual data as input for a depression prediction model, and then applies a saliency-based post hoc analysis to enhance interpretability. In this research the aforementioned work of Othmani, Zeghina, and Muzammel (2022) has been used as a base for the structure of the depression predictor which will be explained in more detail in chapter 3 Rather than focusing solely on predictive accuracy, our approach identifies the most influential temporal segments within the audiovisual data. This enables us to understand which specific moments drive the model's decisions, offering a more transparent and clinically relevant insight into the prediction process.

2.2. Video Analysis

Video processing is a key field within computer vision, allowing the analysis of dynamic content over time. Video processing enables computers to track movements, analyze events or changes in behavior and environment, by breaking down a sequence of video frames. Video processing methods can be applied to a variety of computer vision tasks and it is important especially in applications that require an understanding of temporal relationships in video data Tang et al. 2023. In healthcare, video processing opens up innovative approaches for patient monitoring and diagnosis. Its ability to detect subtle changes in a patient's emotional or physical state through facial expression or body movements has made video-based analysis a powerful and promising tool for gaining insights into conditions like depression, stress or other neurological disorders. There are numerous techniques to extract useful information from video, we will delve in the ones used for the scope of this research.

2.2.1. Visual modality

Facial Features associated with depression

Depression can alter someone's facial emotion and non-verbal behavior in observable ways. Studies have identified several facial cues that are common in individuals with depressive symptoms. Pupil

dilation has been associated with depression. Siegle et al. (2011) suggest that faster pupil movements are associated with healthy controls, whilst depressed subjects present slower pupil dilation responses under some conditions (He et al., 2022). Several studies suggest that an individual diagnosed with depression, displays low expressibility in their facial expressions (He et al., 2022). That includes reduced eye contact (Lucas et al., 2015), gaze direction, eyelid activity (Alghowinem, Goecke, Wagner, et al., 2013), iris movement, and eye openings/blinking (Alghowinem, Goecke, Cohn, et al., 2015). For the scope of this research, a subset of facial features has been used for depression classification. By focusing on these key indicators, we can capture the nuanced ways in which depressive symptoms manifest. For instance, reduced positive emotions such as smiling less often (e.g. infrequent lip-corner pulling smiles) has been associated with depressed individuals (Mahayossanunt et al., 2023). Additionally mentioned in the work of Mahayossanunt et al. (2023), depressed individuals tend to display less overall facial activity and expressivity. Besides the reduced positive emotions, there are several negative affect cues that have been associated with depression. In depression, certain facial expressions associated with sadness or distress occur more often. For example, depressed patients may exhibit more frowning or scowling expressions which convey negative feelings and social disinterest.

To analyze these subtle variations rigorously, we use the Facial Action Coding System (FACS). FACS deconstructs complex facial expressions into individual, measurable action units, allowing us to objectively quantify facial movements. This is especially important for explainability tasks, because by mapping individual muscle movements to specific emotional states, FACS helps researchers and practitioners trace model decisions back to observable, interpretable actions.

Facial Action Coding System (FACS) & Action Units (AUs)

The Facial Action Coding System (FACS) (Ekman and Friesen, 1978) is a framework used to taxonomize human facial movements. FACS classifies all visually distinguishable facial activity on the basis of 44 unique Action Units (AUs), along with categories of head and eye positions and movements. Action Units represent the fundamental actions of individual muscles or groups of muscles and have been tested on offset/onset duration, average duration, occurrence frequency, and offset/onset ratios (Ekman and Rosenberg, 1997).

- **Offset/Onset duration**: This reference to the duration that a muscle takes to begin its movement (onset) and the time it needs to reverse in its neutral state (offset). Testing this helps identify how quickly or slowly certain facial actions occur.
- **Average duration**: Typical duration in which the AU is active during a facial expression. Measuring this gives us insight in how long each facial movement usually last. This gives valuable information for depression prediction tasks, for example Parikh, Sadeghi, and Eskofier 2024 state that people with depression exhibited prolonged facial movements linked to sadness and reduced intensity of happy expressions
- **Occurrence frequency**: It measures how often a particular AU appears in among various facial movements. Measuring this help us determine how common or rare are facial expressions in different contexts.
- **Offset/Onset ratios**: This examines the relationship between the onset and offset of the muscle. By analyzing this, we can determine whether facial expressions change gradually or abruptly.

Table 2.1 and Table 2.2 depict the AUs coded in FACS and the muscle groups involved in each action. In the second figure more grossly defined action units are listed which means more obvious and pronounced movements.

AU Number	Descriptor	Muscular Basis
1	Inner Brow Raiser	Frontalis, Pars Medialis
2	Outer Brow Raiser	Frontalis, Pars Lateralis
4	Brow Lowerer	Depressor Glabellae, Depressor Supercilli, Corrugator
5	Upper Lid Raiser	Levator Palpebrae Superioris
6	Cheek Raiser	Orbicularis Oculi, Pars Orbitalis
7	Lid Tightener	Orbicularis Oculi, Pars Palpebralis
9	Nose Wrinkler	Levator Labii Superioris, Alaeque Nasi
10	Upper Lip Raiser	Levator Labii Superioris, Caput Infraorbitalis
11	Nasolabial Fold Deepener	Zygomatic Minor
12	Lip Corner Puller	Zygomatic Major
13	Cheek Puffer	Caninus
14	Dimpler	Buccinator
15	Lip Corner Depressor	Triangularis
16	Lower Lip Depressor	Depressor Labii
17	Chin Raiser	Mentalis
18	Lip Puckerer	Incisivii Labii Superioris, Incisivii Labii Inferioris
20	Lip Stretcher	Risorius
22	Lip Funneler	Orbicularis Oris
23	Lip Tightener	Orbicularis Oris
24	Lip Pressor	Orbicularis Oris
25	Lips Part	Depressor Labii, or Relaxation of Mentalis or Orbicularis Oris
26	Jaw Drop	Masseter; Temporal and Internal Pterygoid relaxed
27	Mouth Stretch	Pterygoids, Digastric
28	Lip Suck	Orbicularis Oris

Table 2.1: Action Units and their Muscular Basis

AU number	FACS name
19.	Tongue out
21.	Neck Tightener
29.	Jaw Thrust
30.	Jaw Sideways
31.	Jaw Clencher
32.	Lip Bite
33.	Cheek Blow
34.	Cheek Puff
35.	Cheek Suck
36.	Tongue Bulge
37.	Lip Wipe
38.	Nostril Dilator
39.	Nostril Compressor
41.	Lid Droop
42.	Slit
43.	Eyes Closed
44.	Squint
45.	Blink
46.	Wink

Table 2.2: More grossly defined Action Units (AUs) in the FACS

In this research, we used the DAIC-WOZ dataset that contains 20 of those Action Units detected via the OpenFace toolkit. From the comprehensive review by Yasin et al. 2023 it is clear that DAIC-WOZ

dataset is a widely used dataset for depression prediction tasks. To capture the nuanced facial behaviors associated with depression—both in terms of reduced positive expressivity and increased negative affect—the data must offer detailed and reliable annotations of individual action units. The DAIC-WOZ dataset meets these requirements, as they represent a behaviorally and clinically informative subset of the Facial Action Coding System (FACS) commonly used in affective computing and clinical research, enabling a comprehensive analysis that directly leverages the strengths of FACS. The Action Units contained in the DAIC-WOZ dataset are depicted in Table 2.3. Several studies support the claim that the DAIC-WOZ dataset is a valuable resource for clinical affective behavior analysis, particularly in the context of automated depression detection. It provides rich multimodal data—including audio, video, transcriptions, and facial behavior features—that allow researchers to investigate a wide range of behavioral and emotional cues associated with depressive symptoms. For example Akbar et al. 2021 achieved high accuracy detecting depression using AUs extracted from the DAIC-WOZ dataset.

Action Unit	Description	Muscle Movement
AU01	Inner Brow Raiser	Raises the inner part of the eyebrows.
AU02	Outer Brow Raiser	Raises the outer part of the eyebrows.
AU04	Brow Lowerer	Lowers the eyebrows.
AU05	Upper Lid Raiser	Raises the upper eyelids.
AU06	Cheek Raiser	Raises the cheeks (e.g., during a smile).
AU09	Nose Wrinkler	Wrinkles the nose.
AU10	Upper Lip Raiser	Raises the upper lip.
AU12	Lip Corner Puller	Pulls lip corners upward (smiling).
AU14	Dimpler	Tightens the corners of the mouth.
AU15	Lip Corner Depressor	Pulls lip corners downward.
AU17	Chin Raiser	Raises the chin.
AU20	Lip Stretcher	Stretches the lips horizontally.
AU25	Lips Part	Separates the lips (e.g., during speaking).
AU26	Jaw Drop	Opens the mouth (e.g., surprise).
AU23	Lip Tightener	Tightens the lips.
AU28	Lip Suck	Pulls the lips inward.
AU45	Blink	Closes the eyes (blinking).

Table 2.3: Descriptions and muscle movements for the 20 Action Units used in the DAIC-WOZ dataset.

2.2.2. Audio modality

One of the earliest modern works that associated depression with voice is the work of Kraepelin (1921). In his work he defined depressed voice as 'patients speak in a low voice, slowly, hesitatingly, monotonously, sometimes stuttering, whispering, try several times before they bring out a word, become mute in the middle of a sentence'. Indeed in more recent studies shows that speech features covers 38% of a message for affective computing, which is more that the semantic information acquired from the speech Mehrabian 2017. These features, including aspects such as prosody and acoustic properties, can be measured to capture how emotions are expressed in speech. For example, a slower speech rate combined with reduced pitch variation can make a voice sound monotonous, a characteristic frequently associated with depression. Figure 2.1 provides an illustration of a digital audio signal, demonstrating how these measurable attributes underpin affective analysis.

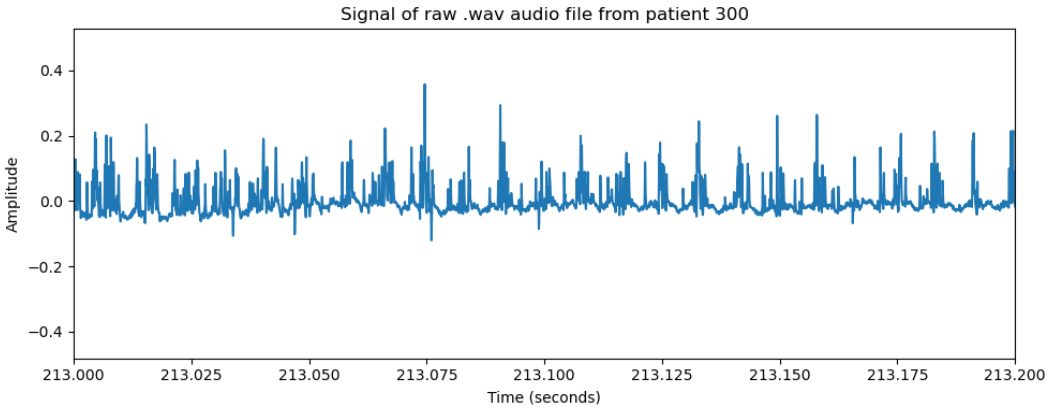


Figure 2.1: Signal of a small time interval of the raw .wav audio file from patient 300

Audio Features associated with depression

Depression also manifests in the acoustic properties of a person's speech. A number of paralinguistic vocal features have been identified as markers of depression. Mundt, Vogel, et al. (2012) state that depressed speech tends to be slower. Additionally, patients speak at a reduced rate and take more frequent or prolonged pauses mid-conversation. In recordings, severely depressed individuals produced significantly longer silence intervals, more variable pause lengths, and a higher proportion of time spent pausing versus speaking. Another mark is a flattening of vocal prosody. Depressed individuals often speak in a monotone, with reduced pitch variation and inflection (Girard and Cohn, 2015). In study by Liang et al. (2024) a set of 331 acoustic features (spanning cepstral, spectral, and voice quality domains) could accurately distinguish individuals with MDD from healthy controls, indicating that a rich acoustic profile of one's voice is highly informative of depression status.

To analyze those variation in the voice, we use Log-Mel spectrograms, which provide a detailed time-frequency representation of speech. Log-Mel spectrograms capture the rich spectral characteristics of speech, making them well-suited for detecting subtle prosodic and articulatory changes associated with depression.

Log-mel spectrograms

Log-mel spectrograms provide a compact and intuitive representation of audio that closely aligns with human auditory perception—an essential consideration for analyzing speech in the context of depression. It begins with capturing the raw audio signal and running it through the Short-Time Fourier Transform (STFT) Equation 2.1, which decomposes the signal into the frequencies it's made up of, and for how long.

$$[H]\text{STFT}\{x[n]\}(m,\omega) = X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \tag{2.1}$$

Where:

- $x[n]$ is the discrete signal.
- $w[n-m]$ is a window function centered at m.
- ω is the angular frequency.
- $X(m,\omega)$ is the STFT result, a function of time (represented by m) and frequency (represented by ω).

However, while the STFT offers precise frequency information, it does not reflect how us humans actually hear sounds. Human auditory perception is non-linear: we are more sensitive to differences in lower frequencies and perceive changes in loudness on a logarithmic scale. To bridge this discrepancy, the frequency components are transformed using the Mel scale—a scale designed to mimic the

non-linear sensitivity of the human ear. This transformation compresses higher frequencies and expands lower frequencies, effectively emphasizing the frequency bands that humans are more in tune to. Figure 2.2 visually demonstrates this conversion from Hertz to Mel.

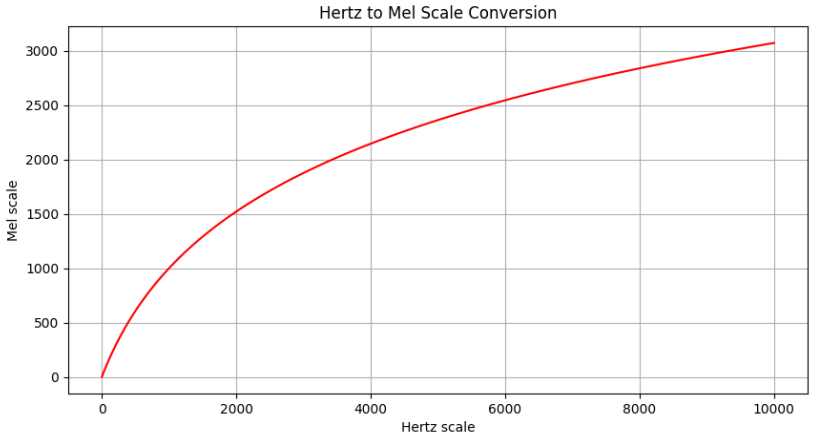


Figure 2.2: Hertz to mel conversion

Building on this, a mel spectrogram is generated by applying a series of Mel filter banks to the STFT output, grouping frequencies into bins that mirror our perceptual resolution. By converting the resulting mel spectrogram's magnitudes to a logarithmic scale, we obtain a log-mel spectrogram that not only preserves the essential frequency details but also mirrors perceptual loudness differences. An example of a log-mel spectrogram is illustrated in Figure 2.3.

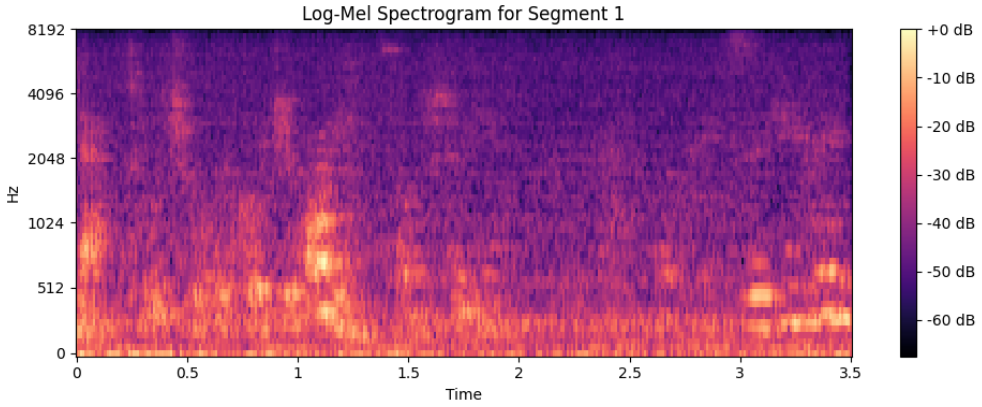


Figure 2.3: Log-mel spectrogram for Segment 1 of Participant 300

In the context of depression, several indicators become more apparent through this logarithmic representation. For instance, as mentioned above, a slower speech rate is common in depressed individuals Mundt, Vogel, et al. (2012). In the context of log-mel spectrograms it manifests as extended segments with minimal variation in spectral content. Similarly, reduced pitch variation results in a narrow range of frequency modulation over time, contributing to a monotonous tone. At the same time, lower overall vocal energy appears as a consistent drop in amplitude across key frequency bands, and subtle shifts in energy distribution, such as reduced high-frequency components, can signal a lack of vocal vibrancy. These interpretable features extracted from log-mel spectrograms provide vital cues for understanding the acoustic correlates of depressive speech, thereby forming a foundational element in machine learning approaches to depression detection.

2.3. Deep learning

2.3.1. Neural networks

A neural network is a structure that is inspired by the structure and function of biological neural networks found in living organisms, such as humans and animals. Neural networks are designed to help machines learn complex patterns and get valuable insights from input data . Given their ability to automatically learn hierarchical features, neural networks have become a key tool in affective computing and mental health research. In this thesis, neural networks play a crucial role in both feature extraction and depression prediction. For audio feature extraction, we utilize AlexNet, a deep convolutional neural network (CNN), to process Log-Mel spectrograms and extract relevant acoustic features. For visual feature extraction, a CNN is employed to capture facial action units and other visual cues indicative of depression. Finally, these extracted features are used as inputs to a simple Multi-layer Perceptron (MLP), which serves as the core depression prediction model. Later we apply a saliency-based post hoc analysis in the depression predictor.

Multi-layer Perceptrons (MPLs)

Multi-layer Perceptrons is a type of feed-forward network that consist of multiple layers of neurons, where each layer is fully-connected to the next one. As mentioned above, a Multi-layer Perceptron is employed to perform the depression classification task. Each node, or neuron, receives multiple inputs which are either outputs from previous layer's nodes or raw data from the input. Each input is multiplied by a weight that represents the importance of the corresponding input to the node. The input layer receives the data and it consists of as many nodes as the number of features in the data. Then one or more hidden layers are responsible for learning more complex representations and extract features from the input data. An example of an MLP is illustrated in Figure 2.4.

To achieve that each node calculates a weighted sum of all its inputs:

$$z_j = \sum_{i=1}^{n} w_{ij} x_i + b_j \tag{2.2}$$

, where,

- z_j is the result of the weighted sum for neuron j in current layer.
- w_{ij} the weight of the connection between current neuron j and previous neuron i,
- x_i are the inputs of neuron j/ output of previous neuron i,
- b_j is the bias term of the current neuron j, and
- n the number of neurons in the previous layer,

After the weighted sum is calculated for each neuron is passed through an activation function. Since the function is non-linear (e.g ReLU, sigmoid) it introduces nonlinearity which enhances the model's capability to learn complex patterns. A linear model would not be able to learn complex relationships in the data since regardless of the number of layers added it would still behave as a linear function. Finally, the output of the activation function will now be either the new input of the next layer of neurons or the output of the model in case of the last layer.

Activation functions

As mentioned above activation functions are crucial for our model's capability of solving non linear problems. An activation function is a differentiable operator that applied fixed mathematical transformation to a node's output. The purpose of the activation function is to decide how much the output contributes to the next layer without involving any trainable parameters. Some of the most popular functions are shown in Figure 2.5.

For our MLP Depression predictor we have used ReLU as an activation function The **Rectified Linear Unit (ReLU)** (Fukushima, 1975) activation function is defined as:

$$f(x) = \max(0, x) \tag{2.3}$$

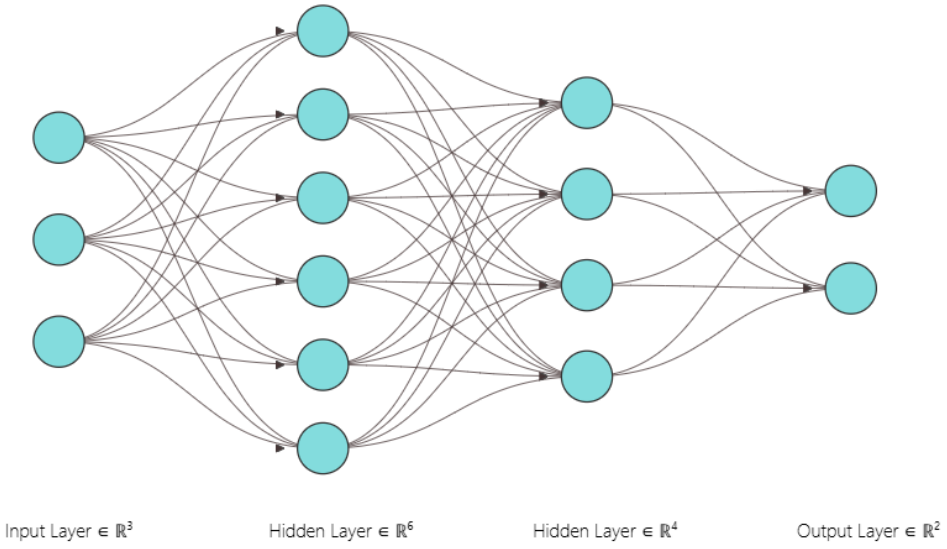
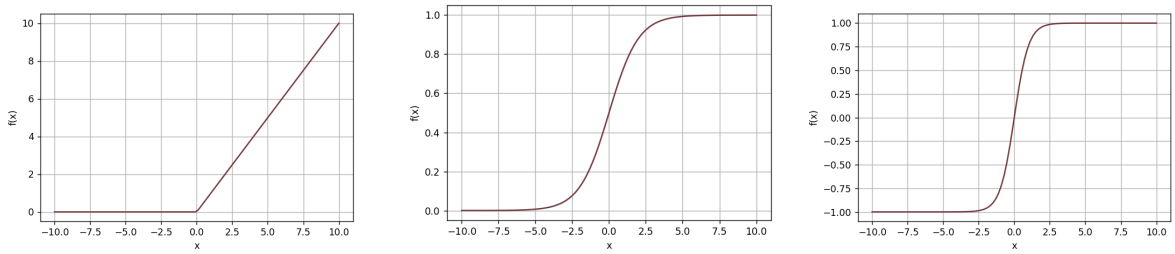


Figure 2.4: A 4-layer MLP with 2 hidden layers



(a) ReLU **(b)** Sigmoid **(c)** Tanh

Figure 2.5: Popular activation functions.

ReLU outputs zero for all negative inputs and the input itself for positive inputs. It is computationally efficient and widely used in deep neural networks.

Additionally, the **Softmax** function transforms the raw output scores of the model into a probability distribution over the classes. In this project, Softmax is applied at the output of the Depression Predictor Model. These probabilities are then utilized in the saliency-based post hoc analysis to interpret the model's decision-making process.

The Softmax function is defined as:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}, \quad \text{for } i = 1, 2, \ldots, n, \tag{2.4}$$

where:

- z_i is the i-th raw output (logit) from the neural network,
- n is the number of classes,
- e^{z_i} represents the exponential function applied to z_i,
- The denominator is the sum of the exponential of all logits, ensuring that the outputs form a valid probability distribution.

Loss functions

Loss functions in general are mathematical functions of the parameters of the machine learning algorithm. They are fundamental components in machine learning algorithms, as they provide a quantitative method to evaluate how well the algorithm is modeling the dataset. A loss function measures the difference between the predicted value and the real target value (Equation 2.5), in order to improve the model's performance. The goal is to minimize the total loss, defined as:

$$\text{Total Loss} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{y}_i) \tag{2.5}$$

where:

- N is the number of data points,
- $L(y_i, \hat{y}_i)$ is the loss for the i-th data point.

The choice of the appropriate loss function L depends on the nature of the task (e.g. classification, regression) and plays a crucial role in the model's training and final results. For example the **Mean Squared Error (MSE)** is a commonly used loss function for regression tasks, while **Cross-Entropy** is used for classification problems.

For the purpose of this thesis we used one modification of the Cross-Entropy Loss, the **Focal Loss**. The Focal Loss is used for tasks where the data is not balanced meaning that one (or more) classes are significantly underrepresented in the dataset. The DAIC-WOZ dataset used for the training of our Depression Predictor is also biased as it contains more participants that are non-depressed than depressed. The Focal Loss works by down-weighting the well represented class, enabling the model to focus more on harder to classify examples. Focal Loss is defined as:

$$L_{\text{Focal}}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}(1-\hat{y}_{ij})^{\gamma}\log(\hat{y}_{ij}) \tag{2.6}$$

where:

- N is the number of data points,
- C is the number of classes,
- y_{ij} is a binary indicator (1 if class j is the correct class for data point i, otherwise 0),
- $\hat{y}_{ij}$ is the predicted probability for class j for data point i,
- γ is the focusing parameter that controls the strength of down-weighting for well-classified examples ($\gamma > 0$).

2.3.2. Convolutional neural networks

A **Convolutional Neural Network (CNN)** (LeCun et al., 1989) is a specific type of deep neural network designed to learn patterns in structured data, including both spatial data (e.g., images) and temporal data (e.g., sequences). The difference from the traditional neural networks is that CNNs use convolutional layers to extract patterns from the data. The architecture typically consists of an input layer, hidden layers and an output layer, while the hidden layers include one or more layers that perform convolutions. While originally developed for image processing, CNNs are also effective for analyzing one-dimensional time-series data.

1D CNN

A 1D Convolutional Neural Network (1D CNN) is a type of convolutional neural network specifically designed to process sequential data rather than spatial data like images. The "1D" in 1D CNN refers to the fact that convolutions are applied along only one dimension—typically the time axis or sequence axis. In this project, we leverage a 1D CNN to extract meaningful temporal features from sequences

of facial Action Units (AUs), enabling the model to learn patterns in how facial muscle activity evolves over time—patterns that may be indicative of depression. Unlike fully connected networks, 1D CNNs can efficiently capture short-term dependencies in sequential data, making them well-suited for this task as shown in the work of Kurek, Świderska, and Szymanowski (2024). In Chapter 3 we provide a detailed explanation of how the 1D CNN extracts high level features from the Action Units and how these features are then fed to the depression predictor. Figure 2.6 provides an overview of the 1D-CNN architecture from Othmani, Zeghina, and Muzammel 2022 which we used for our visual features extraction.

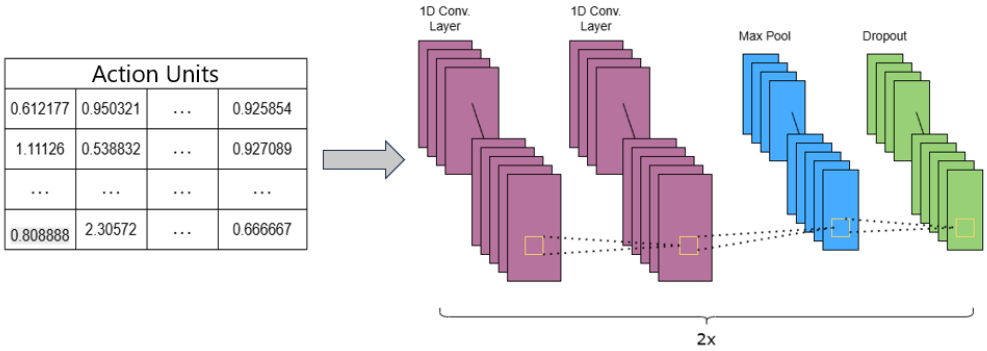


Figure 2.6: 1D CNN

AlexNet

AlexNet, introduced by Krizhevsky, Sutskever, and Hinton 2012, is a Convolutional Neural Network (CNN) architecture that played a significant role in advancing deep learning research. It gained attention after achieving strong performance in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), demonstrating the effectiveness of deep CNNs for large-scale image classification tasks. AlexNet consists of eight layers—five convolutional layers followed by three fully connected layers—that extract hierarchical features, from low-level textures to high-level shapes and objects. The architecture is illustrated in Figure 2.7. While originally designed for image classification, AlexNet has been successfully adapted for audio-based tasks by processing Log-Mel spectrograms—which are time-frequency representations of sound. Since spectrograms share structural similarities with images, CNNs like AlexNet can effectively capture relevant spectral and temporal patterns.

In this thesis, AlexNet is leveraged to extract acoustic features from Log-Mel spectrograms of speech data. These extracted features capture key speech characteristics, such as prosody, pitch variations, harmonic structure, and energy distributions, which have been shown to be indicative of depression. Specifically, AlexNet enables the model to identify depression-related vocal markers, such as monotonic speech, reduced pitch variation, and abnormal prosody—all of which have been linked to Major Depressive Disorder (MDD). The extracted deep spectral features from AlexNet are then fed to our depression predictor, where they are combined with visual features extracted from Action Unit sequences using a 1D CNN as mentioned above. In Chapter 3 we explain in more detail the procedure in which AlexNet extracts those high level features from the Log-mel spectrograms. This multi-modal approach integrates both audio and facial behavioral cues, providing a more comprehensive framework for depression detection

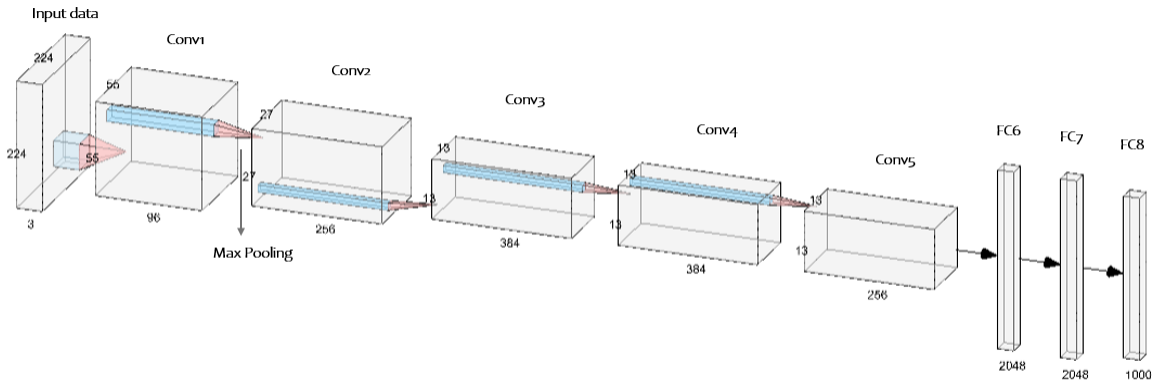


Figure 2.7: AlexNet architecture

Model Calibration

Model calibration is the process of aligning a model's predicted probabilities with the actual likelihood of those predictions. In a well-calibrated model, the predicted confidence aligns with reality. For example, if a model predicts an event with confidence 80%, that event should occur in 80% of similar cases. This calibration is important in applications such as medical diagnosis, risk prediction, and autonomous systems, where accurate uncertainty estimates inform critical decisions. Poorly calibrated models can make over-confident or under-confident predictions, negatively affecting downstream tasks such as classification reliability and model interpretability.

In this thesis, model calibration plays an important role in ensuring the reliability of the depression prediction probabilities, which are later analyzed using a saliency-based explainability approach. Since our saliency method relies on analyzing the gradients of entropy derived from the probability distributions, it is essential that these probabilities accurately reflect true likelihoods rather than being overconfident or poorly distributed. Without proper calibration, the entropy-based saliency scores could be misleading, failing to capture meaningful uncertainty information in the model's decisions.

Temperature scaling is a post processing method for re-calibration of the model, especially neural networks. It works by adjusting the predicted logits (the raw outputs of the model before applying softmax) using a temperature parameter $T > 0$. This adjustment "softens" the probabilities, reducing the overconfidence when $T > 1$. The adjusted probability for a class i is calculated as:

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}, \tag{2.7}$$

where z_i represents the logit for class i. The temperature parameter T is typically learned in a waiting validation set to minimize the calibration error. Because temperature scaling is performed as a post-processing step, it does not require the retraining of the model, and thus is very computationally cheap and simple to implement. This has made it a pretty popular method for deep learning calibration, especially for classification tasks. In this work, temperature scaling is applied to the depression predictor model to calibrate its probability outputs before performing the saliency analysis in those calibrated probability distributions. In the next section, we introduce the saliency-based approach used in this study and explain how entropy gradients are leveraged to interpret the model's decision-making process.

2.4. Explainable AI

Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms. This is especially important in applications where decision-making has significant consequences, such as medical diagnosis, mental health assessment, and risk prediction, where understanding the reasoning behind a model's decisions is just as important as the predictions themselves. Without explainability, deep learning models are often regarded as 'black boxes', making it difficult to assess their reliability and potential biases. To address this, we incorporate a saliency-based XAI method in our depression classification pipeline. The goal is to interpret which aspects of the input —derived from speech and facial Action Units (AUs)—contribute most to the model's decision, offering insight into the model's behavior.

2.4.1. Saliency

Saliency methods are widely used XAI tools that highlight the most influential parts of the input contributing to a model's prediction (Schuff et al., 2022; Hu et al., 2023). In the context of this thesis, saliency refers to the identification of time segments in a patient's audio-visual data that most influenced the model's classification of depression. By highlighting these key features, saliency methods help reveal how neural networks process information, offering valuable insights into their decision-making process. As such, saliency is used as an XAI tool since it helps us understand which parts of the inputs are the most influential in generating the output.

A notable contribution to the field is the work by Raman et al. (2024), who propose an information-theoretic saliency framework for time-series forecasting. Their method identifies the most salient timesteps in an observed sequence by quantifying how much each timestep contributes to reducing the model's uncertainty about future outcomes. By grounding saliency in differential entropy, they link the concept of feature importance to information gain, rather than arbitrary perturbation-based metrics. This approach enables principled counterfactual reasoning in forecasting tasks, and supports human-in-the-loop analysis by highlighting which observations were most informative in shaping the model's probabilistic forecast.

Building on this idea, in this thesis we adopt the entropy-based saliency method tailored to our depression detection model. Instead of using a time-series data, we treat our inputs as time series by applying a sliding window in the input modalities that represent a video. This way we will find how influential each segment was for the model. In Chapter 3 we explain with more detail how the approach by Raman et al. (2024) is applied for our depression predictor's probabilities. As mentioned by Raman et al. (2024), the definition of saliency given by Loog (2011) is:

$$S(x) := \det(J^\top \varphi(x) J\varphi(x)) \tag{2.8}$$

, where J_ϕ denoted the Jacobian matrix of ϕ which is the feature mapping.

To visualize the the amount of influence each input component, in our case segments, have in the model's output we use saliency maps. The mathematical formulation above leverages the determinant of the product of the Jacobian's transpose and the Jacobian itself, effectively measuring the local sensitivity of the feature mapping. This determinant reflects how small perturbations in the input can lead to significant changes in the output, thereby quantifying the "information gain" or uncertainty reduction attributed to each segment. This visualization is particularly valuable in our application, as it not only highlights the critical moments in a patient's audio-visual data but also illustrates how these influences evolve over time. In Figure 5.13 a saliency map is illustrated that represents the saliency of the segments contained in one of the 8.5 second video clips inlcuded in the experiments.

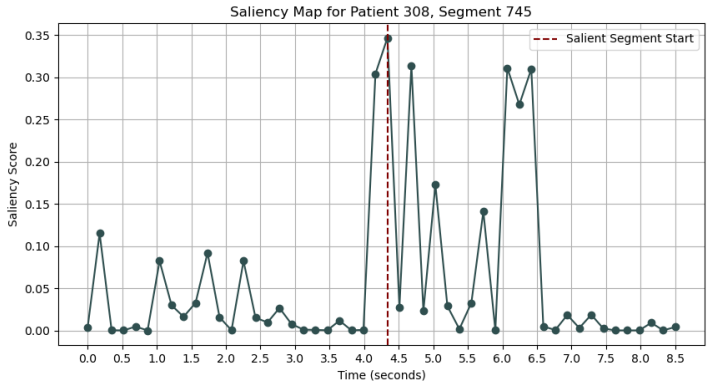


Figure 2.8: Saliency map of Clip 308, 8.5 seconds

2.5. Research Gap

Explainable AI (XAI) has increasingly been applied to depression prediction to not only achieve high predictive accuracy but also to provide transparency in decision-making—a critical factor in clinical contexts. For instance, a recent work by Jo et al. (2024) demonstrates that post hoc XAI methods such as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin, 2016) and Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) can reveal which facial and acoustic features drive predictions in depression models. These techniques assign importance scores to features, so they offer insights into how non-verbal cues - like subtle changes in facial expression, variations in vocal prosody, and shifts in body language - contribute to the classification of depressive symptoms. These methods have been employed for depression prediction tasks to interpret feature contributions at the individual prediction level (Jo et al., 2024; Al Masud et al., 2025). Although informative, these explanations are inherently static (Guidotti et al., 2018). They tell us which cues are important on the whole, but not when during the interaction those cues become important. As a result, such explanations ignore the **temporal evolution** of depression-related behaviors. Recent work highlights that many explainability methods in depression detection still focus on static or frame-level interpretations and fail to adequately incorporate temporal cues, limiting their utility for understanding the evolution of affective behavior over time. For example, Moreno et al. 2023 notes that “interpretability from temporal activities from videos when deep models are used is not fully explored,” highlighting the gap in time-aware explainability approaches.

Our saliency entropy-based approach overcomes the above limitations. In contrast to existing methods that offer static, frame-level explanations, our approach provides a dynamic, time-aware analysis that captures the temporal evolution of depression-related behaviors. A saliency-based explainability approach directly targets this gap by capturing **when** and **where** the model is attending to depression-relevant cues in the input video/audio sequence. Additionally, compared to other temporal approaches such as Attention mechanisms our saliency entropy-based approach offers a model-agnostic, post-hoc method that leverages the temporal relationships of the input (Gomez, Fréour, and Mouchère, 2022). By computing the entropy of the model’s probability distribution and deriving the Jacobian with respect to each video segment, this approach provides a more precise, fine-grained sensitivity analysis that reflects the causal impact of small perturbations in the input. By filling this gap, our method contributes to the broader goal of enhancing interpretability in AI-driven depression detection, and helps us answer the first part of the research question *’To what extent can interpretability in AI-based medical diagnosis for depression be improved by identifying the salient moments in video and audio data and investigating them through human evaluation?’*.

To answer the second part of the research question is also important for interpretability. Most prior works on XAI presents AI-highlighted segments or features in isolation, without rigorously comparing them against human-selected cues or expert annotations (Y. Zhang et al., 2023). Hence, a significant gap lies in the limited effort to systematically validate these explanations against human judgment (Qi et al., 2023). Recent work is beginning to acknowledge and address this gap. K. Yang et al. (2023)

address this by conducting *"strict human evaluations to assess the quality of the generated explanations"*, creating a dataset of human-annotated explanations for comparison. L. Zhang et al. (2025) compare and contrast their model explanations with the professional annotations: they create a test set of *"text chunks identified by human annotators as important for depression prediction"* and show that their system could retrieve similar segments as supporting evidence. Such efforts underscore that verifying AI explanations against expert annotations or user feedback is important. Following the example set by these studies, we conducted our own experiments with lay participants. In our experiments, participants reviewed the model's highlighted audiovisual segments and answered targeted questions regarding the presence and relevance of depression cues. It directly addresses our research question by answering wether investigating and evaluating salient audiovisual moments through human evaluation can potentially enhance the interpretability of AI-based depression diagnosis.

3

Methodology

This research develops a method that integrates salient moment detection with statistical feature analysis to improve the interpretability of depression prediction. Our approach is two-fold. First we implement a model for segment-wise depression prediction and, based on probability distributions, identify salient segments for each patient. Second, we examine these salient segments by employing saliency maps and rigorous statistical analyses for short video clips (containing those salient moments) that were presented to human participants. This section outlines the methodology for developing and evaluating the depression predictor, for extracting the salient moments and for performing the statistical analysis to assess them.

3.1. Data Collection

The initial phase involves collecting a dataset that contains a sufficient amount of individuals who either suffer from depression or not. There are plenty of datasets in the literature either containing text data, EEG data, or audiovisual data. The focus of this research is depression prediction for audiovisual data and this is why we chose the Distress Analysis Interview Corpus (DAIC) dataset Gratch et al. 2014. The DAIC dataset contains clinical interviews that aim to support the diagnosis of mental health conditions associated with distress such as anxiety, depression, and post-traumatic stress disorder. A subset of this dataset, the one used for this project, is called DAIC-WOZ. This dataset focuses solely on depression, as it contains depressed and non depressed individuals. Data collected from DAIC-WOZ dataset include audio recordings and visual related information from interviews and extensive questionnaire responses. The visual related information contain Action Units and also facial features representing movements of specific points in the face per timestamp. The interviews are conducted by an animated virtual agent, Ellie, which is controlled by a human interviewer in another room.

3.2. Data Preprocessing

Following a preprocessing procedure similar to the work of Othmani, Zeghina, and Muzammel (2022), the speech recording and the visual features are aligned and then divided into smaller segments of 3.5 seconds. The audio and visual data are truncated and synchronized using the start and end timestamps of the transcript, which include the full interview text along with detailed speaker annotations and timestamps. Deviating from the work by Othmani, Zeghina, and Muzammel (2022), which just segments the data into 3.5 second segments, we decided to follow a sliding window setup. A sliding window framework is employed to localize salient moments with finer temporal resolution, allowing us to identify critical segments within shorter time intervals. The setup that was chosen was a window size of 3.5 seconds with a stride of 0.1 seconds. This configuration was chosen to balance two critical factors. First, a 3.5-second window is short enough that the incremental addition of 0.1 seconds still influences the segment's content, thereby ensuring that the calculated changes in entropy reflect meaningful shifts in the signal. If the window were considerably larger, the 0.1-second shift would represent a very small fraction of the total segment, reducing its impact on the overall entropy and potentially

masking subtle yet important variations. At the same time 3.5 seconds are long enough to contain meaningful information for the model to learn to classify depression and it was used in the work of Othmani, Zeghina, and Muzammel (2022) which was the work that part of our preprocessing procedure was based. Second, the stride of 0.1 seconds is a good compromise for the trade-off temporal resolution/computational cost. While a smaller stride could further increase temporal resolution, it would also generate an excessive number of highly overlapping segments, leading to increased computational costs without proportionate gains in the detection of perceptually relevant changes.

Additionally, using the transcript data, only the audio data where the participant speaks is kept. This means that all the audio where the interviewer speaks is zeroed out. We kept the visual features for those moments because even if the patient does not speak, useful visual cues can still be captured in the patient while the interviewer is speaking. The next step, to prevent overfitting in the deep neural networks, an extra preprocessing step is applied in audio recordings. First, the raw audio is augmented to increase data variability. In this step, a bit of noise is introduced by randomly selecting a sample from the audio and subtracting a percentage (between 1% and 10%) of its amplitude from the entire signal. This perturbation is to simulate natural background variations. The perturbed audio is then further augmented by randomly shifting its pitch by up to ±2 semitones. This pitch augmentation is employed to enhance the diversity of the training data so that the network can generalize more effectively to new samples. Finally, the AUs file contain Action Units and the changes in their values over the timestamps. Gratch et al. (2014) created this file using OpenFace (Baltrušaitis, Robinson, and Morency, 2016), where from the raw videos they extracted the Action Units. There are some parts of the videos where the OpenFace algorithm was not confident in its detection of facial features. In practice, these values suggest that the face was either not detected properly—perhaps due to occlusion, poor lighting, or an angle that makes detection difficult—or that the tracking failed. Consequently, the data for those frames is not reliable and does not contain any useful information about the facial movements. Those frames were removed from the data along with the respective frames of the audio data, so the data stays synchronized.

3.2.1. Audio data preprocessing

After the aforementioned preprocessing in the audio data, the calculation of log-mel spectrograms follows, which is a type of visual representation of audio signals. First, we define the parameters for the short-time Fourier transform (STFT): a window length of 25 milliseconds and a hop length of 10 milliseconds, which are calculated based on the audio sampling rate. The STFT is then performed using a Hann window, and the squared magnitude of its complex values is taken to generate the power spectrum. This power spectrum is converted to the Mel scale using 64 Mel bands, a transformation that aligns the frequency representation with the human auditory system by emphasizing perceptually important frequency bands. Finally, the Mel-spectrogram is converted to a logarithmic scale through a decibel transformation, which compresses the dynamic range of the data and prepares it for further analysis or model training, while ensuring numerical stability by adding an offset to avoid taking the logarithm of zero.

Extracting log-mel spectrograms is used as feature extraction since they can capture essential features of phonemes and intonation, which are critical for speech recognition systems. In the Othmani, Zeghina, and Muzammel (2022) paper, the obtained log-mel spectrograms are passed through a VGGish network (Hershey et al., 2017) for extraction of high-level features. Within our high-level audio feature extraction approach, we leverage a similar deep neural network architecture, an AlexNet model, as defined in Chapter 2. To be able to do that, we reshape these 2D spectrograms into a pseudo-3D representation. This is done by replicating the spectrogram into three channels and enhancing it with temporal derivatives, the delta and delta–delta features. The output is a three-channel representation that preserves the spectral information but is compatible with architectures like AlexNet. We then employ a modified version of AlexNet inspired by the work of Venkataramanan and Rajamohan (2019). AlexNet was originally designed for three-channel RGB images; rather than altering its architecture to accept a single-channel input, we repeat the log-mel spectrogram into three channels. In **ModifiedAlexNet**, convolutional layers extract spatial features in the spectrogram, while the classifier, which is a dropout layer followed by a linear layer, projects these features into a small representation (with a default of 512). The changes to the network, such as resizing the sizes of the convolutional kernels and strides to match the sizes of the spectrograms, are drawn from previous research on emotion detection

(Venkataramanan and Rajamohan, 2019) and are designed to better capture the salient audio features that accompany depression. For every patient, each segment of the log-mel data is transformed into a three-channel format. The segments are then individually passed through the modified AlexNet model to obtain high-level feature representations. These audio -features are aggregated for each patient and they will be one of the two modalities that will be fed as an input to train the segment-wise Depression Predictor. Figure 3.1 depicts the audio preprocessing procedure that was described above.

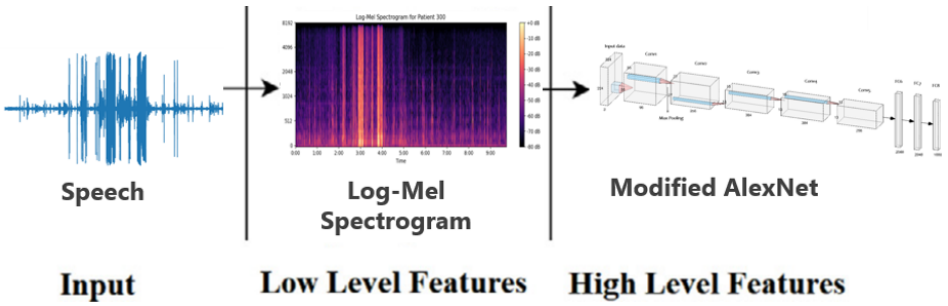


Figure 3.1: Audio preprocessing diagram

3.2.2. Visual data preprocessing

As mentioned before, the DAIC-WOZ dataset does not contain the raw videos of the interviewees for confidentiality reasons. Therefore, our analysis relies on pre-extracted visual features the Action Units (AUs) instead. The AUs are used as visual features to encode the facial patterns of depressed subjects. As described in Chapter 2, AUs capture subtle facial muscle movements according to the Facial Action Coding System (Ekman and Friesen, 1978) and serve as indicators for differentiating between depressed and non-depressed individuals. Using a 1D-CNN is particularly appropriate as convolutional layers are good at capturing local patterns in sequential data, such as trends and temporal dependencies, that are crucial to model the evolution of facial expressions over time. Inspired by the configuration in Othmani, Zeghina, and Muzammel (2022), our AU1DCNN model is configured to receive the 20-channel input encoding the 20 AUs. It employs a number of convolutional layers with large kernel sizes and matching padding to collect a sufficiently large temporal window, followed by pooling and dropout layers to reduce dimensionality and prevent overfitting. In addition, we incorporate a dimensionality reduction module—implemented as a fully connected layer (FeatureReducer)—to project the high-level features generated by the CNN into a lower-dimensional space to match the dimensionality of the audio modality. Again each segment of AU data is independently fed to the 1D-CNN. The extracted visual features are the second of the two modalities that are fed as an input to train the segment-wise Depression Predictor. A visual representation of the procedure described above can be found in Figure 2.6 in Chapter 2.

3.3. Model Development

In our approach, multi-modal features from the audio and vision modalities are combined to obtain a robust depression predictor. We initially as mentioned above extract and preprocess features: audio features are extracted from log-mel spectrograms and reduced in dimensions, and visual features are extracted from Action Unit readings that capture facial expressions. These features are then concatenated per segment to generate a combined multimodal representation of each segment in an interview.

Since every patient's data is divided into multiple segments under the sliding window framework, we perform a flattening operation that concatenates all segment-level features from every patient into a single tensor for training. This means that our training procedure is completely patient independent, since the segments are flattened and then shuffled before they are fed into the model. This flattened information, together with depression labels, is used to create a custom PyTorch Dataset. The dataset also keeps track of patient IDs and segment indices, which can later be used to reconstruct the original order of the segments within each patient for the saliency calculation.

The depression predictor model itself, is a deep neural network with several fully connected layers

interspersed with batch normalization, ReLU activations, and dropout layers. The network takes the high-dimensional multimodal feature vectors as input and maps them to a lower-dimensional space before a final classification into two classes (depressed or non-depressed). This architecture is designed to successfully reduce the complexity of the input features without sacrificing the most salient information necessary for prediction. The Adam optimizer is used with weight decay for regularization and a focal loss function is used as the training criterion as an attempt to mitigate the class imbalance that favors the non-depressed group. The focal loss down-weights easy examples and focuses the model on harder, misclassified examples, which is crucial in scenarios where one class might be underrepresented. In addition, we incorporate learning rate scheduling to reduce the learning rate when the training loss plateaus.

We use a softmax function for the model's output, which produces probability distributions per segment for calculating saliency. Calibration is essential when using softmax as the output of the neural network, to ensure that the predicted probabilities are meaningful and they accurately reflect the real probabilities of the event. Calibration ensures that when the model assigns a certain probability to a prediction—say, 70%—this value accurately reflects the true likelihood of depression. In our calibration step we employ temperature scaling to adjust the model's probability estimates. After scaling the model, we compute the softmax probabilities from the logits and store these alongside the corresponding true labels and patient/segment identifiers that will later be used for the reconstruction.

To evaluate how well our model's predicted probabilities align with actual outcomes, we generated reliability diagrams, shown in Figure 3.2 and Figure 3.3. A reliability diagram is a visual tool for assessing how well a model's predicted probabilities match the actual outcomes. Before Calibration (Figure 3.2): The blue line is considerably away from the diagonal, indicating that the model's probability estimates are not well calibrated. There are bins in which the model overestimates and in which it underestimates the true probability of depression. After calibration (Figure 3.3), the curve of the calibrated model is more aligned with the diagonal. This suggests that for a predicted probability p to a prediction, that probability is now a better approximation of the true fraction of depressed cases in that bin.

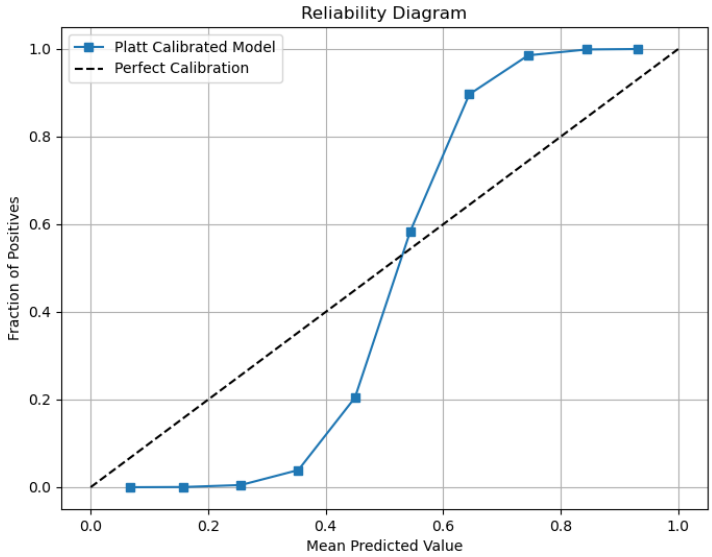


Figure 3.2: Reliability diagram before calibration

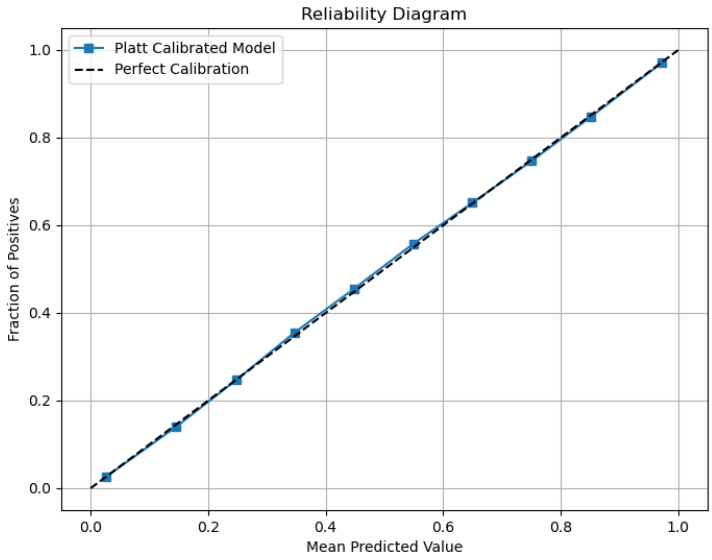


Figure 3.3: Reliability diagram after calibration

3.4. Saliency

To further interpret our model's decisions, we compute saliency maps over the segments for each patient. Building on the methodology proposed by Raman et al. (2024) for using saliency maps to explain why the model forecasted a future, saliency here is defined as the degree of change in the model's entropy of predicted probabilities from one segment to the next, capturing how *informative* each segment is for the final classification.

We begin by loading calibration results, which contain each segment's predicted probability distribution, the ID of the patient that the segment belongs to, and the segments' original index within each patient. The probability distributions were generated by the calibrated model. We then sort and group segments by patient ID, ensuring that each patient's segments are processed in chronological order based on their original segment indices. To compute the saliencies, the entropies need to be calculated first. Since the output of the predictor is a discrete distribution, the entropy is calculated by the above definition:

$$h(X) = -\sum_i p(x_i) \log p(x_i) \tag{3.1}$$

where X is the predicted probability distribution for that segment. Higher entropy indicates greater uncertainty in the model's prediction.

Having obtained entropy values for each segment of a patient, we then find the discrete gradient of the values across consecutive segments. In one dimension, such a gradient is simply the Jacobian of the entropy with respect to the index of the segment. The Jacobian in this case quantifies the sensitivity of the entropy (and hence of the model uncertainty) to infinitesimal changes in time. By squaring the Jacobian (gradient) values, we get our measure of saliency based on Equation 2.8. A large squared gradient indicates a steep transition in entropy between adjacent segments, highlighting a segment where the model confidence changes significantly and thus suggesting that the segment is salient.

Because raw saliency scores can vary from patient to patient, we normalize them to the range of [0,1] by subtracting the minimum and then dividing by the maximum minus the minimum. We then identify the top 5 segments with the highest saliency for each patient. From those highest saliency segments we will later pick a number of them to include in the experiments.

3.5. Evaluation

The first step of the evaluation is the evaluation of the model's performance. The model's performance is evaluated with a list of quantitative metrics, primarily through a classification report and monitoring the training loss over epochs. The classification report provides key performance metrics like accuracy, precision, recall, and F1 scores for depressed and non-depressed classes, providing a general overview of the ability of the model to classify between these two classes. In this study, model evaluation is performed exclusively on the training data. We monitor the training loss over epochs to assess how steadily the model learns and converges; a consistently decreasing training loss indicates that the model is capturing meaningful patterns from the data. Although relying solely on training loss and training-based classification reports may overestimate performance due to potential overfitting, these metrics offer valuable insight into the model's learning dynamics and its ability to correctly classify cases during the development phase. Future work will extend the evaluation framework aiming to improve model generalization in unseen data.

Using the procedure described in Section 3.4, salient segments are calculated. The salient segments though represent a window of 3.5 seconds. Since we work in a sliding window framework with a stride of 0.1 seconds there is a substantial overlap between consecutive segments. That means that if a segment S is identified as salient, it means the model's confidence changed significantly relative to segment S-1. However, since S and S−1 share most of their frames (they differ by only 0.1 seconds at the start of S-1 or end of S), the triggering factor for saliency is likely the newly added 0.1-second portion or the 0.1-second portion that was removed from the beginning of segment S−1. Figure 3.4 provides a clear illustration of the above setup. For the purpose of this thesis, we will focus on the added information, hence in the second point of interest.

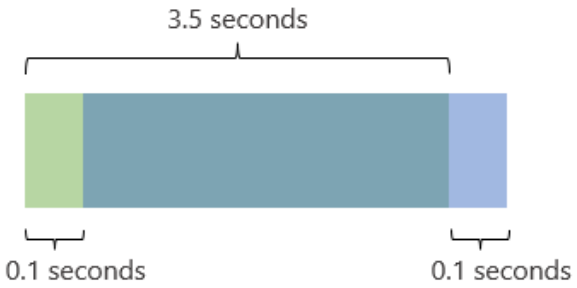


Figure 3.4: Segments S_{-1} (green) and S (blue)

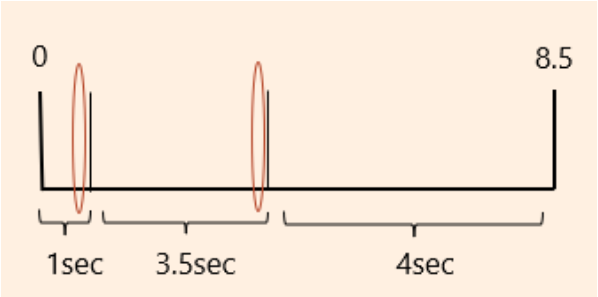
The reasoning behind the focus on the second point of interest is firstly we want to give emphasis on emerging cues. That is because participants are more likely to intuitively focus on information that is *added* rather than removed to make a decision. Study by Yantis and Jonides (1984) supports the idea that our visual attention system is designed to prioritize sudden changed in the environment, suggesting that our attention system is sensitive to new appearing stimuli more than information that simple ceases or is removed. Similarly, auditory research shows comparable effect, for example Näätänen et al. (2007) using the mismatch negativity (MMN) response indicates that our auditory system automatically detects and prioritizes new or deviant sounds.

These points of interest is the reason why, in this research, we distinguish between salient segments—the short, automatically identified intervals derived from a sliding-window analysis—and salient moments, which are the specific points of interest within those segments. Each video clip is approximately 8.5 seconds long, which means it contains 51 consecutive segments. To avoid position bias around the clips the segments are not positioned identically across all the clips. There are 3 different cases regarding the position of the 'salient segment' in the video clip.

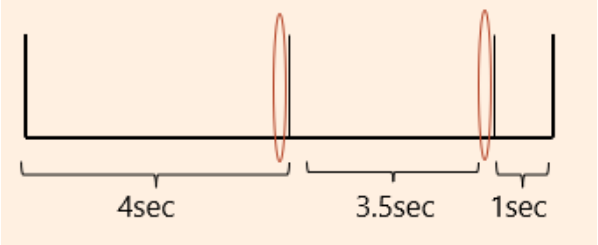
- Case 1: The salient segment is positioned at the beginning of the clip. This means that there is 1 second before the beginning of the 3.5 second segment and 4 seconds after Figure 3.5a.

- Case 2: The salient segment is positioned at the end of the clip. This means that there are 4 seconds before the beginning of the 3.5 seconds segment and then 1 second after Figure 3.5b.
- Case 3: The salient segment is positioned at the middle of the clip. This means that there are 2.5 seconds before the beginning of the 3.5 seconds segment and then 2.5 seconds after Figure 3.5c.

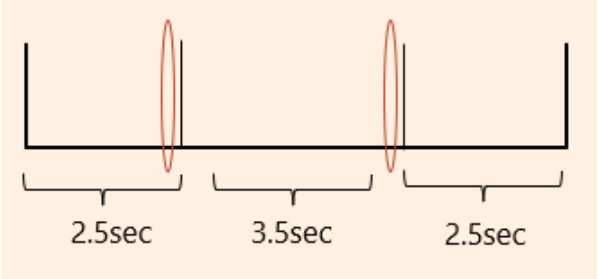
In the figures below we can see the 3 position cases and the points of interest in Figures 3.5a, 3.5b, 3.5c.



(a) Case 1: the segment labeled as salient occurs at the beginning of the 8.5 second video clip



(b) Case 2: the segment labeled as salient occurs at the end of the 8.5 second video clip



(c) Case 3: the segment labeled as salient occurs in the middle of the 8.5 second video clip

Figure 3.5: Three cases based on 3.5 salient segment position

Following the identification of these salient moments, an **experiment** (described in Chapter 4) presents the 8.5-second clips to participants in a questionnaire. Participants are presented with the clips and asked questions about the moments they assess to be salient, how they interpret the behaviors or cues displayed, and whether they agree with the model's identified moments. This setup allows us to investigate:

- In how far the model detects significant intervals that can trigger depression behaviors
- How interpretable or self-explanatory these moments are to human audiences
- Understanding which facial and vocal characteristics are critical for identifying depression—and whether these cues can be reliably recognized by human observers.

By combining quantitative performance indicators (accuracy, sensitivity, specificity) against qualitative

human judgment on salient events, we gain both technical and user-level feedback on the capability of the model to recognize and interpret potential signals of depression within short video segments.

3.6. Statistical Methods

Linear/Logistic mixed effect model

Linear mixed effects models extend traditional regression by incorporating both fixed effects (factors we are explicitly interested in) and random effects (sources of random variability). This method is particularly well-suited for data with a hierarchical or nested structure—where multiple observations come from the same participant or the same clip—because it accounts for the non-independence of these observations. In our study, the fixed effect examines the relationship between depression status and confidence ratings. The random effects capture the variability across participants and clips, ensuring that the effect of depression is estimated accurately without being confounded by these other sources of variability. In our initial model, we included random intercepts for both Participant ID and Clip ID to account for variability at both levels. However, including Participant ID resulted in a singular fit warning, meaning that the variance component for participants was estimated to be nearly zero. This indicates that participants, on average, did not differ substantially in their baseline confidence ratings. Given the limited number of participants (17) and the minimal variability among them, including this term did not improve model fit and only added unnecessary complexity. Consequently, we simplified the model by retaining only the random intercept for Clip ID, which captures the variability in confidence ratings attributable to differences among clips. We implemented this analysis using the lmerTest package in R, which builds upon the lme4 framework. While lme4 is widely used for fitting LMMs, it does not provide p-values for fixed effects by default. lmerTest addresses this limitation by using methods like Satterthwaite's approximation to compute p-values, thereby enhancing the inferential power of our model. This approach allowed us to robustly test whether depression status significantly predicts confidence ratings, while appropriately controlling for both participant-level and clip-level variability in our dataset.

In addition to linear mixed-effects models for continuous outcomes such as confidence ratings, we also employed logistic mixed-effects models to analyze binary classification outcomes (e.g., whether a participant correctly identified a depressed clip). These models are a form of generalized linear mixed models (GLMMs), which extend the mixed-effects framework to handle non-continuous response variables. Specifically, we modeled the binary variable CorrectPrediction (1 = correct classification, 0 = incorrect) using a logit link function, appropriate for binomial outcomes.

Pearson Correlation Analysis

This statistical method measures the linear relationship between two continuous variables. In our study, we calculate the Pearson correlation coefficient between the model's saliency values and the density of participant-selected timestamps for each video clip. For each clip, saliency values provide a measure of where the confidence of the model is most variable. At the same time, the density of participant-labeled timestamps indicates in which locations human evaluators focus their attention when asked to label salient points. By computing the Pearson correlation coefficient, r, between these two variables using the formula:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}},$$

where x_i and y_i denote the corresponding saliency value and corresponding human timestamp densities, we have a quantitative estimate of how well human ratings correlate with those derived from modeling. A high positive correlation indicates that timestamps with high saliency are also frequently selected by human evaluators, suggesting a strong alignment between the automated and human interpretation of the data.

4

Experiments

In the previous chapter we discussed the methodology for detecting salient segments for each patient. Those segments represent moment of high saliency value which indicated that there was an activity in the video that 'surprised' the model and influenced its decision process. This could be anything related to the trainable features such as facial movement and changes in the voice prosody. The goal of this experiment evaluate those salient moment based on human perception and investigate the reasons behind their saliency, aiming to answer the research question posed in Chapter 1. The experiment will be in a questionnaire form that contains only multiple choice questions. The experiment is separated in two sections each one serving a different purpose and answers different sub-questions of the Research Question. The participants were be shown a number of short video clips and were asked to reply in a series of questions regarding those clips.

4.1. Video-Clip Processing

As mentioned in Chapter 3, we already have the top 5 salient segments of each patient. The next step is to make those segments illustrative to the participants of the experiment so that they can evaluate and investigate those segments. For this purpose we will use the 'features' file given from the dataset for each patient. In this file contains the coordinates of 68 two-dimensional points of the face and how they change for each timestamp. Using those files we recreate the face of the participants being interviewed with an animated face. At the same time we used the .wav recording files for the audio of the clip and we synchronized it with the facial movements using the timestamps that are present in both files. The result is an animated speaking face that resembles the actual interview of the patient. The illustration of this face can be seen in Figure 4.1. Each one of the dots represent a facial point that is contained in the 'features' file and its movement progresses over time as the timestamp increases. The animated video clips were created using HTML and JavaScript based on feature and wav files that have been truncated around the salient segment in three different ways, which will be analyzed below. Besides watching the clip in the HTML page, the function of auto-saving the clip locally is also available.

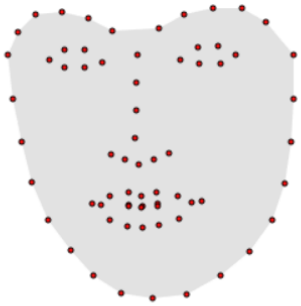


Figure 4.1: Animated speaking face

From the salient segment index (the position of the salient segment in the arrays), we calculate the timestamp taking into consideration one important parameter. In the preprocessing step, some segments that were not valid were removed. Those segments, as mentioned in chapter 3 had one invalid modality which is the video, since in the AUs file for some specific timestamps there is no facial data because the confidence is low. This skews the indexes of the salient segments from their real positions. For this reason a reliability mask has been used to remap all the salient segment's indices with their original indices in the data. After this is done the timestamps of those indices are calculated which they now represent the real timestamp on the interview that this segment represents.

4.2. Experiment-Section 1: General diagnostic question and selected influential/salient moment observations

The first section of the experimental design focuses on answering the first two sub-questions of my research question, "*To what extend can humans identify depression from short video clips?*" and "*To what extend do human-identified salient moments align with the model's salient moments?*". For the first section of the experiment the participants are shown 6 individual video-clips. The first question would be a diagnostic judgment, where participants rate how likely they believe the person in each clip experiences depression. Participant's are asked to assess the depression on a scale of 1 to 10 with 1 meaning they confidence is low that the person experiences depression and 10 confidence is high.

The seconds question the participants are asked to answer is regarding the salient moment identification, where participants will select which moment in the clip was more influential for their decision. This aims to investigate the alignment of the human-identified salient moments compared to the model's selected salient moment. The participants will have a slider that they will slide to the timestamp they chose as influential. Finally, in this section the participants are asked to choose which facial and auditory cues they observed during the influential moment they selected.

4.2.1. Facial cues

The facial cues provided in the multiple choice questions are derived from a set of Action Units that are included in the data set. The AU files contain 20 Action Units which were preprocessed and subsequently used to train our neural network. Those Action Units represent some facial movements, and those movements were offered as multiple choices to the participants. To assist participants, the facial movements are organized in three distinct categories which are the Eyebrows & Forehead area, the Eyes area and the Mouth area. Below are presented choices that were given to the participants and the Action Units that they represent:

Facial Expressions – Eyebrows & Forehead

- **Inner brow raiser (AU01)** → Eyebrows lifted near the center

- **Outer brow raiser (AU02)** → Outer edges of eyebrows raised
- **Furrowed brows (AU04)** → Brow lowering, sign of concentration or distress

Facial Expressions – Eyes

- **Upper lid raiser (AU05)** → Eyes opened wider than usual
- **Eye blink (AU45)** → Repeated blinking, linked to tension or alertness

Facial Expressions – Mouth & Lips

- **Smiling (AU12)** → Lip corner puller
- **Dimpler (AU14)** → Lip corner dimpler
- **Downturned mouth (AU15)** → Indicator of sadness
- **Jaw drop (AU26)** → Open mouth, surprise
- **Lip tightener (AU23_c)** → Pressed lips, sometimes indicating anger or control
- **Lip suck (AU28_c)** → Lips sucked in, sign of nervousness or self-restraint

Parikh, Sadeghi, and Eskofier (2024) mention that depressed patients exhibited higher frequencies of AU1 (inner brow raiser), AU4 (brow lowerer), and AU15 (lip corner depressor), which are commonly associated with sadness and distress. Also, the study found that these patients displayed lower frequencies of AU12 (lip corner puller), which is related to expressions of happiness. There is no strong evidence that AU02 (outer brow raiser) is specifically linked with depression, on the other side is a key component of surprise and fear expressions. Notably, depressed individuals often exhibit an increased blink rate (AU45) compared to healthy individuals and this increased blink frequency tends to normalize as their condition improves (Mackintosh, R. Kumar, and Kitamura, 1983). In summary, excessive blinking (AU45) might indicate internal stress; in depression it has been observed as a physiological change, possibly related to tension or reduced dopamine activity, while in general it's a sign of anxiety or mental strain (Karson, 1988; Kojima et al., 2002). It is widely documented that genuine smiles are less frequent in depressed individuals. Thus, depressed individuals show lower frequency and intensity of AU12 (Smiling) (Sharma et al., 2024). When combined with AU06 (cheek raiser), it produces a "Duchenne smile," indicating genuine happiness (K. M. Sheldon, Corcoran, and M. Sheldon, 2021). AU14 involves tightening the lip corners (often creating dimples); it often appears as a smirk when it is one-sided. Sharma et al. (2024) also suggest that depressed patients exhibit AU14 less frequently than non depressed groups. AU15 (mouth corners pulled downward) is one of the facial actions most associated with depression. Parikh, Sadeghi, and Eskofier (2024) report higher occurrences of AU15 in depressed patients and the action unit is directly linked to expressions of sadness, grief, or despair.

4.2.2. Auditory cues

The DAIC-WOZ dataset provides us the raw audio recordings from the participants. The selection of the multiple choice options for tone of voice was guided by findings from the literature. Key vocal markers – prosody (intonation and pitch variation), speech rate (speed and pausing), loudness (volume/energy), and voice quality measures like jitter and shimmer – have all been examined as potential indicators of depressive states. According to Mundt, Snyder, et al. (2007), patients that responded to depression treatment developed significantly greater pitch variability, paused less when speaking and spoke faster than the baseline, whilst patients that did not respond to treatment did not show similar changes. At the same time Y. Yang, Fairbairn, and Cohn (2012) found out that naive listeners (untrained people) could with satisfying accuracy perceive depression severity from prosody alone. This implies that the intonation changes in depression are salient, supporting the idea that a flat and monotone voice can be a. important sign of depression. Several studies comparing depressed individuals with not depressed group (Taguchi et al., 2018; J. Wang et al., 2019; Shin et al., 2021) suggest that the depressive group had lower voices than the control group, that the tone of voice becomes simpler, lifeless and lower in volume. Jitter and shimmer are also positively correlated with depression (D. M. Low, Bentley, and Ghosh, 2020), indicating a less stable, more hoarse-sounding voice for depressed individuals.

Tone of voice

- **Monotone or flat voice** → associated with depression
- **Slow speech with long pauses** → associated with depression
- **Speech that sounds hesitant or unsure** → associated with depression
- **Reduced loudness** → associated with depression
- **Increased vocal jitter and shimmer** → associated with depression and anxiety
- **Stable voice** → opposite of vocal jitter and shimmer
- **Speech with energy or enthusiasm** → opposite of monotone and flat tone
- **Laughter or lighthearted tone** → not associated with depression
- **Speech that is fluid and uninterrupted** → opposite of hesitant and unsure voice

4.3. Experiment-Section 2: Model's labeled salient moment observations

In this section we are aiming to address the third sub-question, "*What facial/voice features do human identify in AI-selected salient moments for depression?*". The focus on the previous section was on the moments that the participants chose as influential for their decision. In this section we shift the focus on the moments that the model finds influential for its decisions. We are aiming to investigate what do people observe during those moments and how this differentiates from their observations in their own influential moments. Again similarly to section 1 the participants will be show the exact same 6 video clips only now they will be prompted to a specific timestamp. This timestamp will represent the second point of interest which contains the 0.1 seconds that were added due to the sliding window framework. The participants now will be asked to answer the same questions regarding Facial Expressions and Tone of Voice that they were asked before only now they will have to reply with what they observed in the timestamp that they were given.

5

Results

In this section, we present the findings from two complementary perspectives: first, the model's performance on detecting salient moments for depression, and second, the participant experiment evaluating those moments. We begin by outlining the key measures and outputs of the model, demonstrating how it identifies potentially important segments in the video clips. We then discuss the human evaluation, where the observers rated their degree of confidence in picking up on depression, chose their influential moment and selected what facial or vocal cues they believed were having an impact. By considering both the model predictions and the results of the experiment side by side, the goal is to answer as extensively as possible the research question and the sub-questions.

5.1. Model's results

Since the goal is to find patterns of decision making of the model, it is very important that the model's performance in identifying depression is adequate for the task. After training our neural network to identify depression from 3.5 second segments, we evaluated its performance using precision, recall, F1-score, and accuracy. Table 5.1 summarizes the metrics on the training set for each class (0 = not depressed, 1 = depressed), along with macro and weighted averages. We report the model's performance on the training data rather than on new data because the model exhibits a significant class imbalance towards the non-depressed class (Class 0). This class imbalance biases the model's predictions, and therefore it is less accurate for generalizing to unseen data. However, our primary goal here is not to deploy a highly generalized depression predictor; but rather examining the hidden patterns that the model learns in discriminating between depressed and non-depressed. Looking at the training set—where the model actually built its decision boundaries—is how we get closer to unmasking what cues or informative moments it considers relevant. This is significant in regards to understanding the model's internal thought process and where it directs its processing of the data it was trained on, even if its overall performance might not translate so neatly to new, balanced data.

	Precision	Recall	F1-score	Support
Class 0	0.91	0.89	0.90	981323
Class 1	0.79	0.77	0.77	439079
Accuracy			0.86	1420402
Macro Avg	0.83	0.84	0.84	1420402
Weighted Avg	0.88	0.86	0.88	1420402

Table 5.1: Model performance metrics on the training set.

These metrics indicate that the model achieves a relatively high precision and recall for class 0 (not depressed), while class 1 (depressed) shows moderate precision and recall, suggesting some room for improvement. The overall accuracy is 0.86, which is promising given the complexity of the task.

Training Behavior

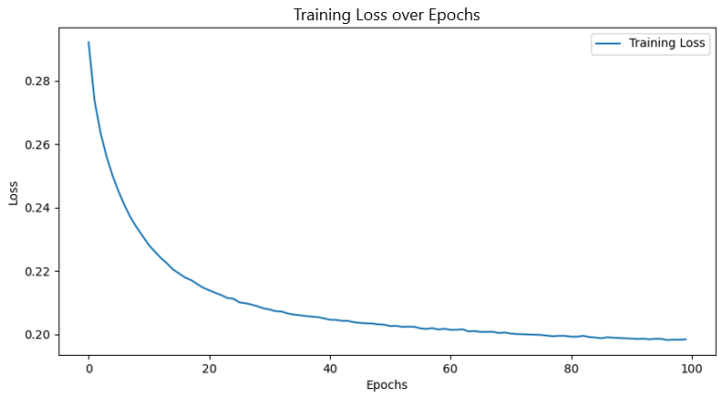


Figure 5.1: Training loss

Figure 5.1 shows the training loss over 100 epochs, illustrating that the model steadily converges. The figure shows a smooth downward trend from approximately 0.28 to about 0.20. This consistent downward slope strongly suggests that the model is successfully capturing patterns in the training data: as the network sees the examples again and adjusts its parameters, it becomes better at minimizing errors on the training set.

5.2. Experiment findings

In this section, we present the results of our human subject experiment for evaluating and interpreting the salient segments of our model. We asked subjects to rate their level of confidence for detecting depression, select the timestamp they thought were most influential for their decision, and specify the facial or voice cues that they observed both in this timestamp and the model's highest saliency moment. By analyzing both their confidence scores and feature selections, we gain insights into how humans perceive and classify the same video clips that our model processed.

5.2.1. SQ1: To what extent can humans identify depression from short video clips?

Classification prediction

Understanding whether human evaluators can accurately identify depression from small video clips is critical to both evaluating the model's outputs and improving its interpretability. To answer this question we need to determine how often participants can correctly identify depression (or non-depression). This includes analyzing overall accuracy, error patterns and confidence level in their judgments.

Figure 5.2 provides a quick visual overview of how often participants correctly identified the clips. By illustrating these proportions in a donut chart, we can immediately see the overall success rate of human judgments. This is a straightforward baseline for how well the participants could tell apart depressed from non-depressed individuals, and setting the stage for more advanced analyses—such as what exactly they were looking at or which moments were more influential. From the illustration below we can see that **66.7%** of the clips where classified correctly.

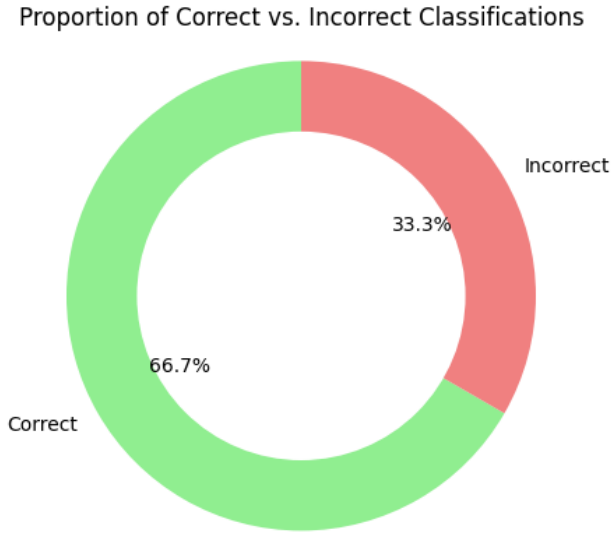


Figure 5.2: Classification prediction

The following table Table 5.2 shows the confusion matrix where the rows represent the actual class and the columns represent the predicted class:

Actual / Predicted	0	1
0	35 (TN)	15 (FP)
1	19 (FN)	33 (TP)

Table 5.2: Confusion Matrix

From this matrix, we calculate the following metrics:

- **Accuracy:**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{31 + 35}{31 + 35 + 17 + 19} \approx 66.7\%$$

- **Precision:**

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{31}{31 + 17} \approx 68.8\%$$

- **Recall (Sensitivity):**

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{31}{31 + 19} \approx 63.5\%$$

- **Specificity:**

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{35}{35 + 17} \approx 70\%$$

The data show that 35 non-depressed individuals were correctly identified, and 33 depressed individuals were correctly classified. In contrast, errors occurred in both directions: 15 non-depressed subjects were mistakenly identified as depressed (false positives) and 19 depressed subjects were overlooked (false negatives). When humans label a subject as depressed, they are correct about 69% of the time (precision), and they detect roughly 64% of all truly depressed cases (recall). These findings suggest

that, although human judgment captures a substantial portion of depression cases from short video clips, there remains a significant margin of error—with a comparable likelihood of missing cases of depression as of over-diagnosing them.

Inter rater reliability

Inter-rater reliability for human judgments of depression presence in the 30 short clips was quantified using Gwet's AC_1, yielding a coefficient of 0.399, which, according to conventional benchmarks, indicates fair agreement among raters but does not reach short of moderate or strong consensus. Each clip was independently rated by either 3 or 4 participants, and the AC_1 statistic accounts for chance-agreement in this unbalanced design. Thus, while raters agreed on depressed versus non-depressed status more often than chance alone, the level of consistency remained modest. However, individually, no rater was below 50% classification rate, with most raters having an above chance score.

Confidence analysis

We want to investigate whether clips labeled "depressed" received different confidence ratings compared to non-depressed clips. In Figure 5.3 we visualize the distribution of participants' confidence ratings (ranging from 1 to 10) for clips that were actually non-depressed (FALSE) versus depressed (TRUE). Each "violin" depicts the full spread and density of confidence ratings: wider sections indicate a higher concentration of ratings at that level, while narrower sections indicate fewer ratings. The boxplots overlaid on the violins highlight the median (thick horizontal line), interquartile range (box), and overall range (whiskers). Notably, the median confidence rating for depressed clips is higher than for non-depressed clips, although there is substantial overlap between the two distributions.

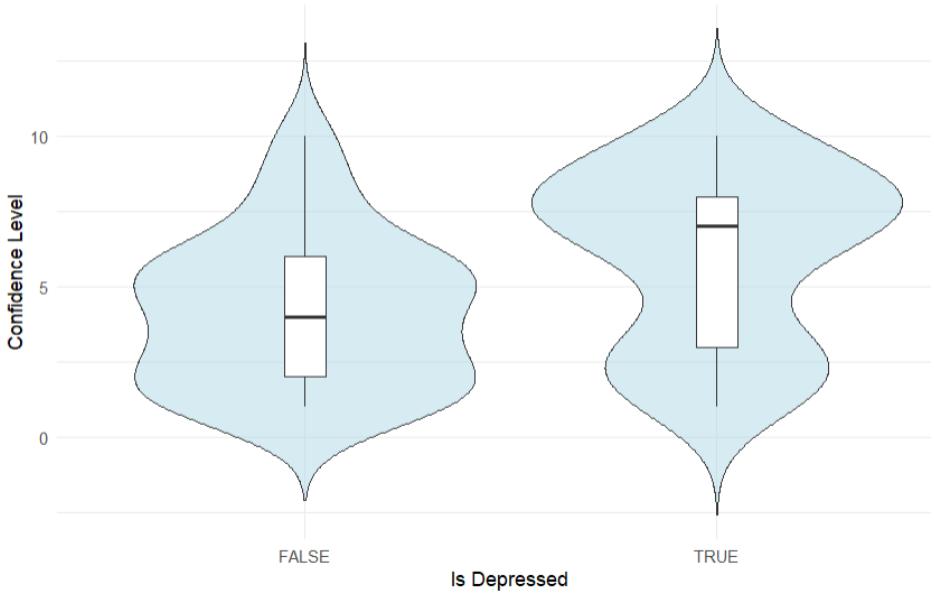


Figure 5.3: Confidence distribution per Class (Depressed vs Non-depressed)

A **Linear mixed-effects model** was fit to quantify this difference and to examine whether this difference is statistically significant or it is by chance. The model (fit by REML) yielded an estimated intercept of 4.11 (SE = 0.59), indicating that non-depressed clips received an average confidence rating of about 4.11. Depressed clips were rated, on average, 1.78 points higher (SE = 0.71) than non-depressed clips, a difference that was statistically significant (t(28.69) = 2.52, p = 0.017). Initially, random intercepts were included for both participants and clips, but the participant-level variance was effectively zero—indicated by a singular fit—so only the clip-level random effect was retained, which had a variance of approximately 1.89.

Additionally, to further deepen out analysis we investigate whether partcipants are more confident when they correctly identify depression (TP) compared to when they incorrectly identify depression. Fig-

ure 5.4 shows that the distribution for TPs is centered at a higher confidence level than for FPs. In line with this pattern, a *Linear mixed-effects model* revealed that participants' confidence was higher for TPs compared to FPs (p = 0.0337, p<.05), with an estimated difference of about 0.95 points in confidence.

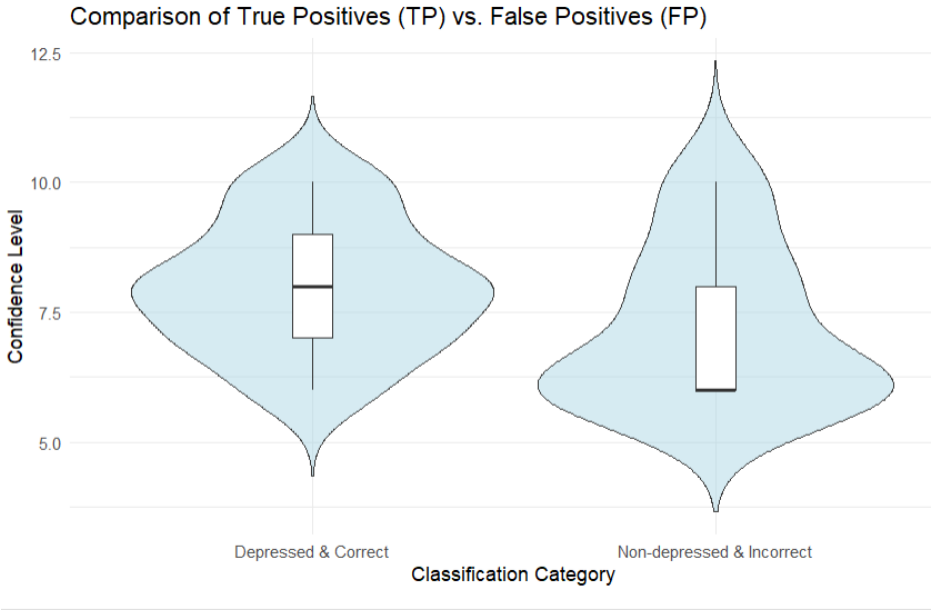


Figure 5.4: Confidence comparison between True Positives and False Positives

5.2.2. SQ2: To what extent do human-identified salient moments align with the model's salient moments?

Alignment with salient moments

One of the key objectives of our research is to find out if the periods that humans identify as most impactful for their decision align with those impactful to our model. If responders repeatedly select time frames that are similar or very close to the model's significant sections, it indicates the AI is picking up on cues analogous to those that humans intuitively use, which could be a promising indicator of interpretability in its decision-making. Conversely, significant discrepancies could indicate that the model focuses on subtleties unrecognized by humans, or that participants notice cues the model overlooks.

In Figure 5.5, we visualize the alignment between human-identified salient moments (represented by violin plots showing the distribution of participants' responses) and model-identified salient moments, for clips where participants correctly classified individuals as depressed (True Positives at the clip level). Each clip's violin plot illustrates how participant-selected moments are distributed within the 8.5-second video clip. The green circles indicate the model's salient segments that were correctly identified by the model (True Positive, TP), whereas red triangles indicate segments where the model misclassified a segment as non-depressed (False Negative, FN).

The figure reveals that participant-selected salient moments generally cluster more closely around the model-identified points when the model correctly classified the segment as depressed (green circles, TP). Conversely, for segments that the model misclassified (red triangles, FN), participant-selected points exhibit a broader distribution, indicating less alignment with the model's salient segment.

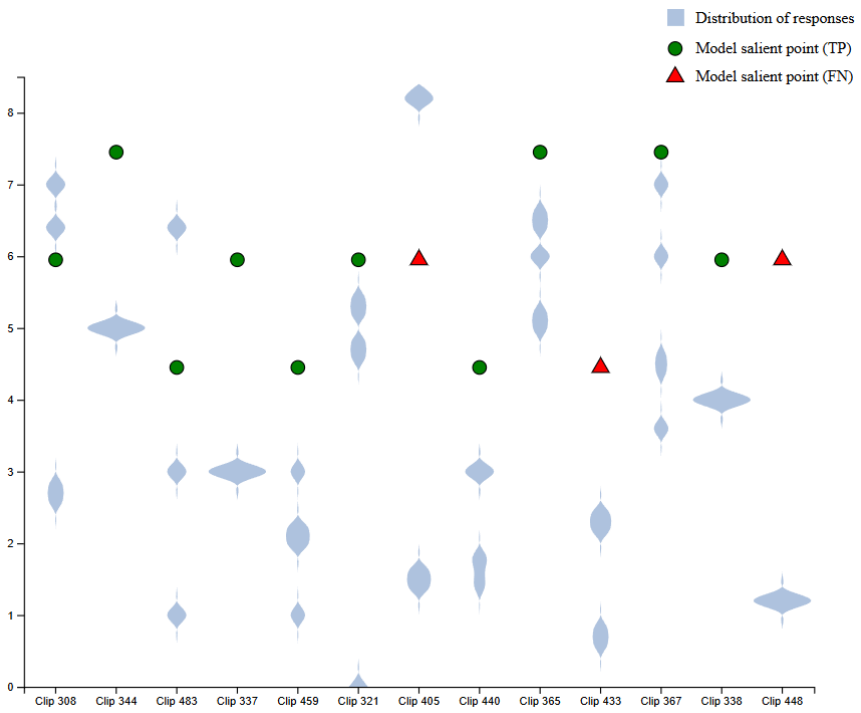


Figure 5.5: Plot for TP classification from participants

In Figure 5.6, we visualize human-identified salient moments for clips where participants correctly classified individuals as non-depressed (True Negatives at the clip level). The violin plots represent the distribution of participant-chosen timestamps for salient segments within each clip. Green circles represent model-identified salient segments that were correctly classified by the model as non-depressed (True Negative, TN), whereas red triangles represent segments incorrectly classified by the model as depressed (False Positive, FP).

Participants' timestamps generally cluster closer to the model's correctly identified salient segments (TN, green circles), although variability remains evident. In contrast, segments classified incorrectly by the model as depressed (FP, red triangles) often show greater variability in participants' selections, indicating less consistency between human-chosen moments and the model's salient segments.

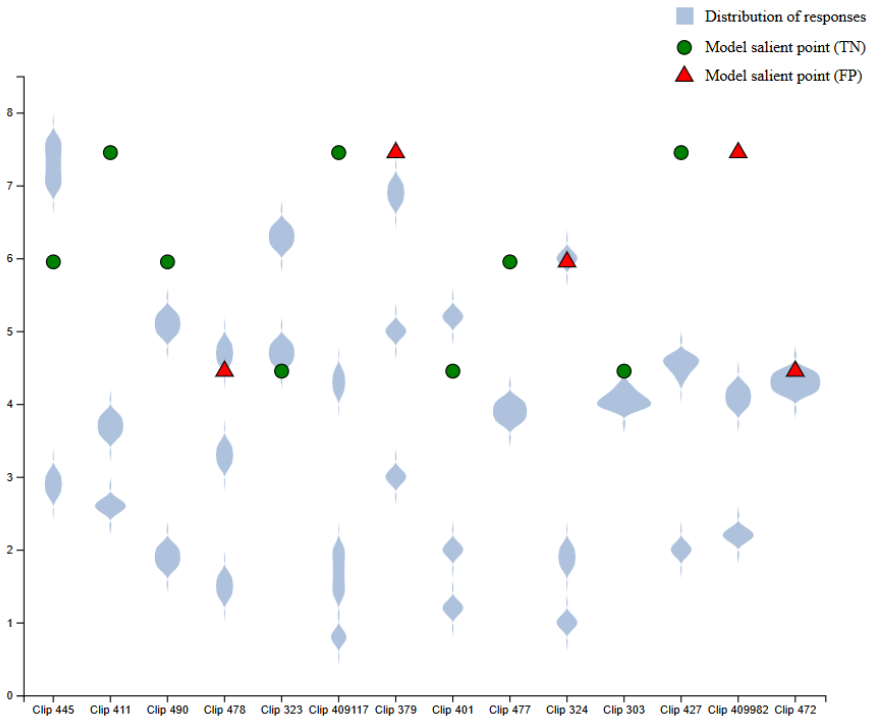


Figure 5.6: Plot for TN classification from participants

In Figure 5.7, we illustrate participant-chosen salient moments for clips where participants mistakenly classified individuals as depressed (False Positives at the clip level). Violin plots represent the distribution of the timestamps selected by participants as salient within each clip. Green circles indicate segments the model correctly classified as non-depressed (True Negative, TN), whereas red triangles indicate segments incorrectly classified by the model as depressed (False Positive, FP). Participant responses appear more dispersed across these clips, regardless of the model's segment classification (TN or FP).

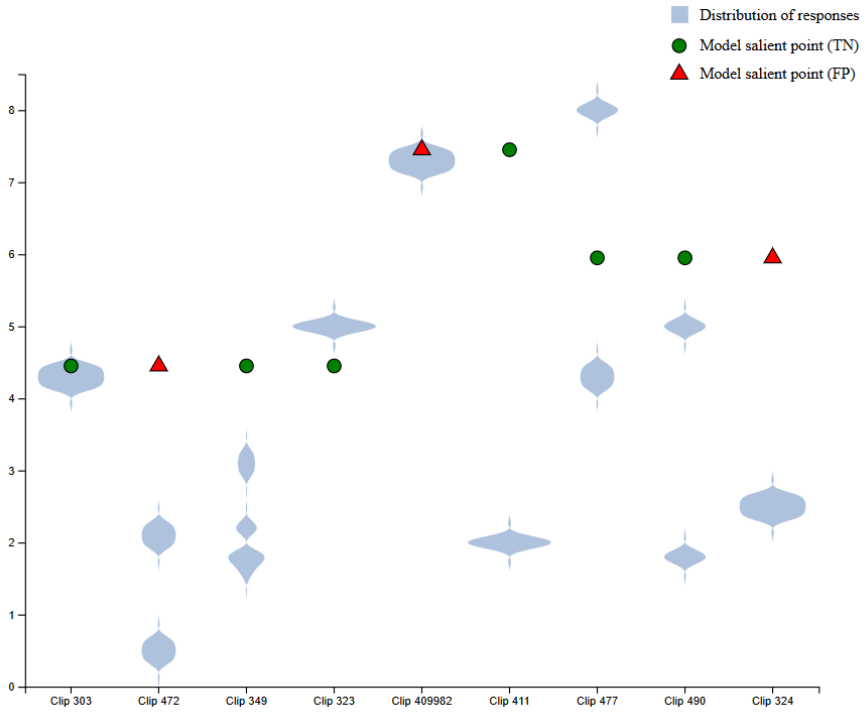


Figure 5.7: Plot for FP classification from participants

In Figure 5.8, we present participants' timestamp choices for clips they classified incorrectly as non-depressed (False Negatives at the participant level). Green circles indicate segments the model correctly identified as depressed (True Positive, TP), while red triangles represent segments incorrectly identified by the model as non-depressed (False Negative, FN).

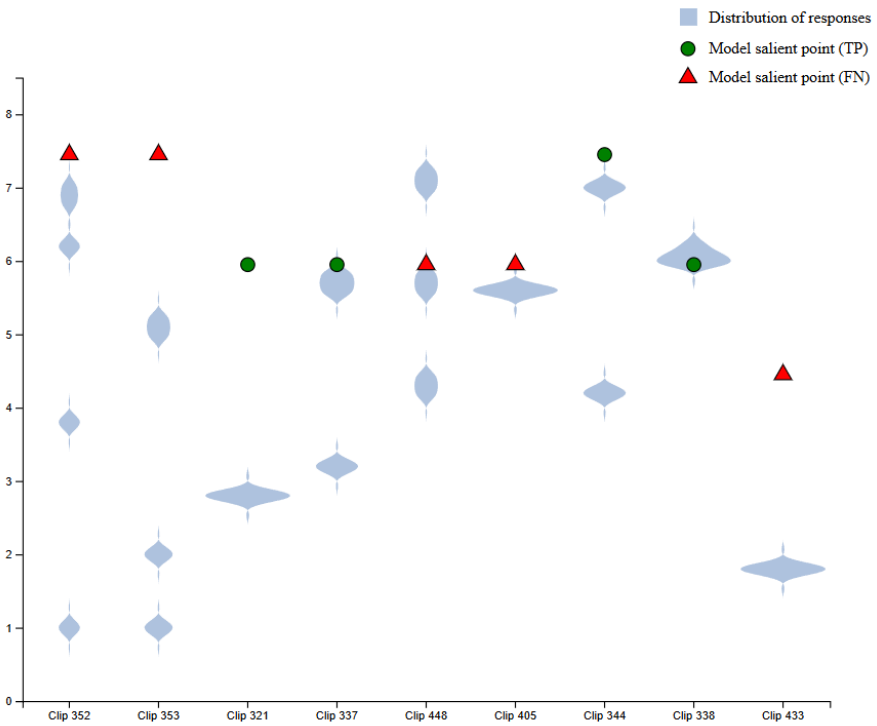


Figure 5.8: Plot for FN classification from participants

Due to simplicity and also due to notable visual differences observed in participant alignment with model-salient moments, our subsequent statistical analysis primarily focuses on clips where participants correctly classified the presence (TP) or absence (TN) of depression. Specifically, we compared alignment for salient segments where the model classification matched participants' correct assessments (TP-TP and TN-TN) versus segments where the model misclassified the salient segment (TP-FN and TN-FP). We excluded detailed analyses of scenarios where participants themselves misclassified clips (FN and FP) because participant uncertainty and variability were inherently higher, limiting meaningful interpretation of human-model alignment.

The above observations are more clear in the illustrations below. In Figure 5.9 and Figure 5.10, we visualize the raw error in seconds between the participants' chosen salient moments and the model's salient segments. Specifically, these graphs show the distribution of errors in the subset of clips where participants correctly identified the presence of depression.

Again a *linear mixed effects model* was conducted to examine whether participants' temporal accuracy—as measured by the raw error between their selected timestamp and the model-derived mean salient interval—differed according to the clip salient moment's classification by the model among cases where participants correctly identified depression. In this analysis, the data were restricted to clips where participants were correct (TP) and were further divided by the model's classification of the salient segment of the clip into two groups: TP (true positive) and FN (false negative). The model included random intercepts for Participant ID and Clip ID. The intercept (–2.48 seconds, SE = 0.73, p = 0.00195) reflects the average raw error for the reference group (clips with FN salient segment classification by the model). Hence on average, the FN group's selected timestamps are about 2.48 seconds before the model's midpoint. The fixed effect for ClipGroupTP was estimated at 0.57 (SE = 0.80, t = 0.71, p = 0.48318), indicating that, on average, the raw error for TP clips was approximately 1.91 seconds ($2.48 + 0.57$) before the model's midpoint. However, this 0.57-second difference does not reach statistical significance (p = 0.48), meaning we cannot conclude that there is a reliable difference between the two groups' Raw Errors.

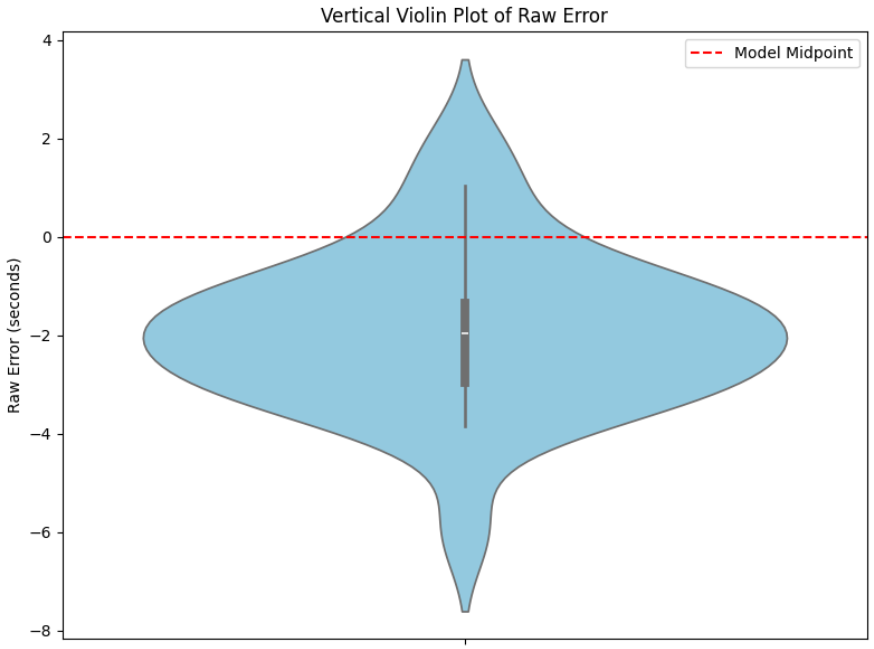


Figure 5.9: Raw error violin plot for TP-TP case

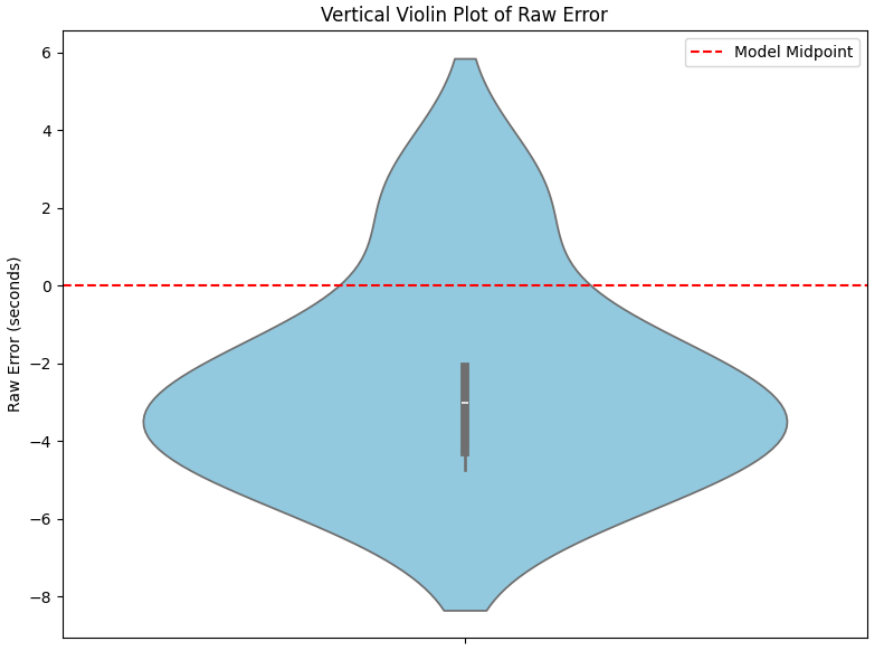


Figure 5.10: Raw error violin plot for TP-FN case

Once more a *linear mixed effects model* was conducted to examine whether participants' temporal accuracy—as measured by the raw error between their selected timestamp and the model□derived mean salient interval—differed according to the clip salient moment's classification by the model among cases where participants correctly identified non-depressed individuals. In this analysis, only cases where participants correctly indicated non-depression (i.e., Confidence Level ≤ 5 and Classification-Correct == "Correct") were included, and the variable ClipGroupND was defined based on the model's classification (TN vs. FP). The model included random intercepts for both Participant ID and Clip ID. The intercept was estimated at –2.36 seconds (SE = 0.89, t(12.36) = –2.63, p = 0.021), representing the predicted raw error for the reference group. The fixed effect comparing TN to FP yielded an estimate of –0.06 seconds (SE = 1.11, t(12.17) = –0.06, p = 0.957), indicating that there was no statistically significant difference in temporal error between TN and FP cases of salient segment in the clips among the correctly classified non-depressed cases. Notably, while there was variability attributable to Clip ID (variance = 2.48), the variability among participants was negligible. The above variability can be seen in Figure 5.11 and Figure 5.12.

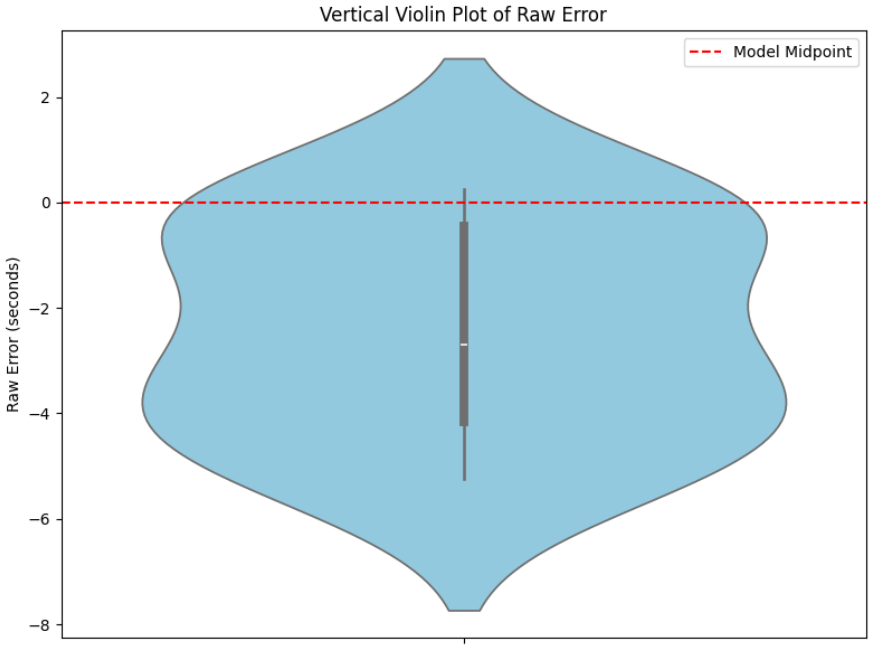


Figure 5.12: Enter Caption

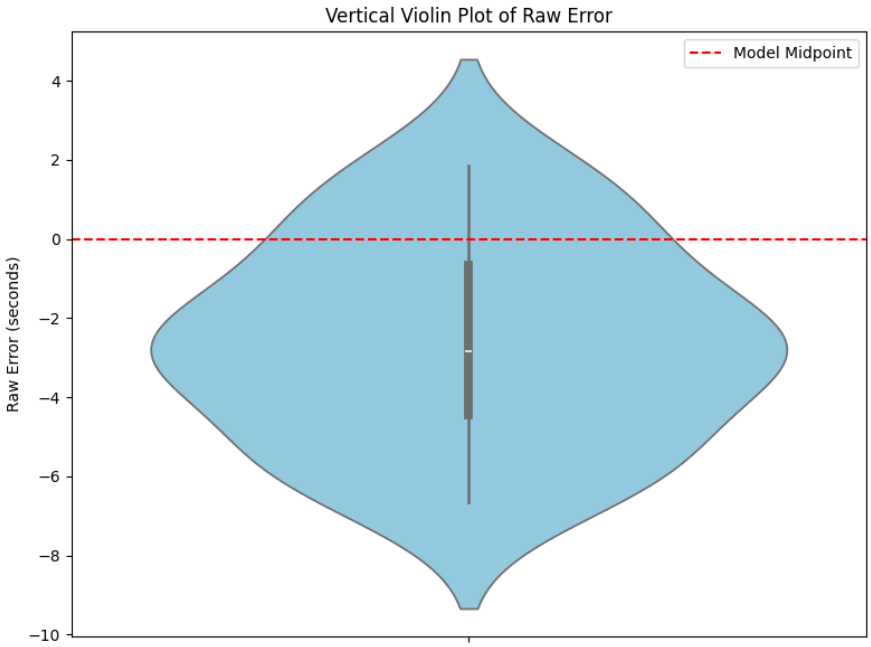


Figure 5.11: Enter Caption

Alignment with saliency maps

Alignment with the salient moment that belongs to the segment which was chosen, as discussed in Chapter 3 as one of the top 5 saliency segments of each patient, is not adequate to determine the overall alignment of saliency between the model and the participants. Therefore, in this section we will explore a more general alignment using the saliency maps of the whole 8.5 second video clips. Figure 5.13 illustrates the model's saliency map for a specific clip (Patient 459), with time in seconds on the x-axis and saliency scores on the y-axis. The dashed red line denotes a segment the model

considers highly salient. By examining those graphs we can observe whether participants lean towards high-saliency regions overall or prefer to choose timestamps in lower-saliency regions often or have no specific pattern. (These observations reveal how closely human judgments align with the model's saliency distribution, whether individuals' cues agree with the model or highly disagree). While the Figure 5.13 depicted the raw saliency curve, we applied a cubic spline interpolation here to produce a smoother line that more clearly highlights the model's major peaks and valleys. The results are depicted in Figure 5.14. Additionally, each red dot indicates a participant's chosen timestamp, allowing us to see whether their selections cluster around higher-saliency moments or diverge from the model's presumed areas of interest.

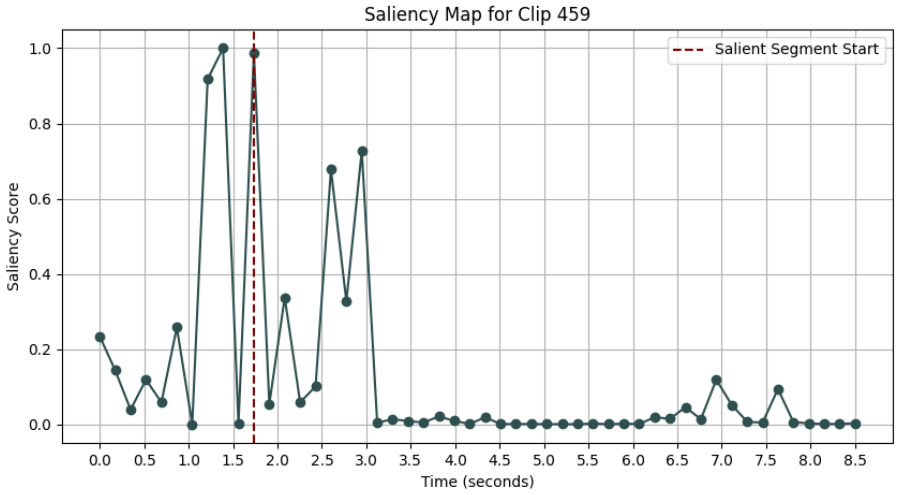


Figure 5.13: Saliency map for 8.5 video clip for participant 459

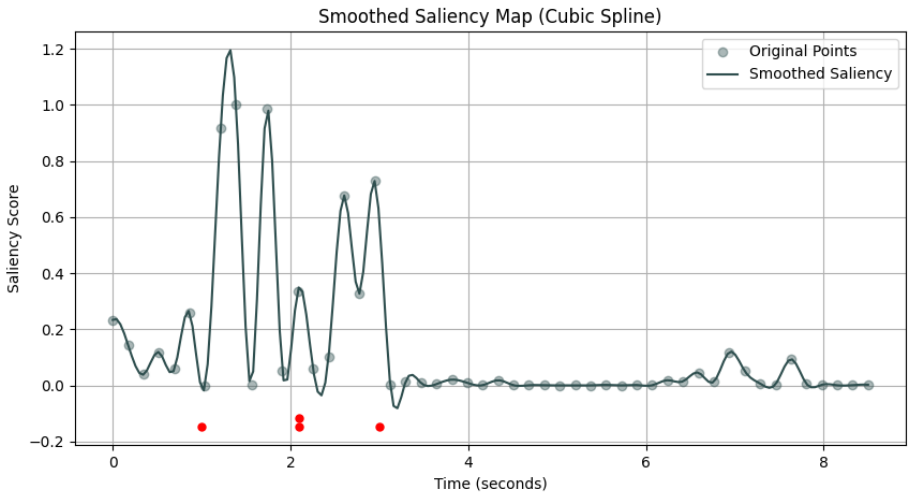


Figure 5.14: Selected timestamps along with saliency map

The figures above illustrate the saliencies of all the segments contained in the 8.5 second video clip that were correctly classifies as depressed from the Depression Predictor (True Positive). As a first step we will investigate the overall alignment of the participants chosen influential models when they correctly classified depressed clips compared to the saliency maps that contain the correctly classified as depressed segments by the model.

Figure 5.15 plots the Pearson correlation (y-axis) for each clip (x-axis) between the model's saliency distribution of TP segments and the density of participants' chosen timestamps when they correctly classified depression. A positive correlation would mean that as the saliency is higher the timestamp density is higher which means that participants choose timestamps with higher saliencies. By examining correlations for all clips that were correctly classified as depressed from the participants, we can see whether participants consistently fall within high-saliency intervals or whether their decisions vary from the model's presumed region of interest. Overall, the correlations vary substantially across the clips. Several clips exhibit positive correlations (e.g., Clip 337, Clip 403, and Clip 405), indicating that participants may have chosen timestamps that tend to coincide with segments the model deemed highly salient for depression detection. Conversely, other clips show near-zero or even negative correlations (e.g., Clip 308, Clip 321, and Clip 344), suggesting that in those clips, participants may have selected timestamps that do not align, or may even diverge, from the model's high-saliency regions. However, none of the above correlation has a p-value below the conventional significance threshold (0.05) indicating that the observed correlations could be due to random chance rather than a true underlying relationship.

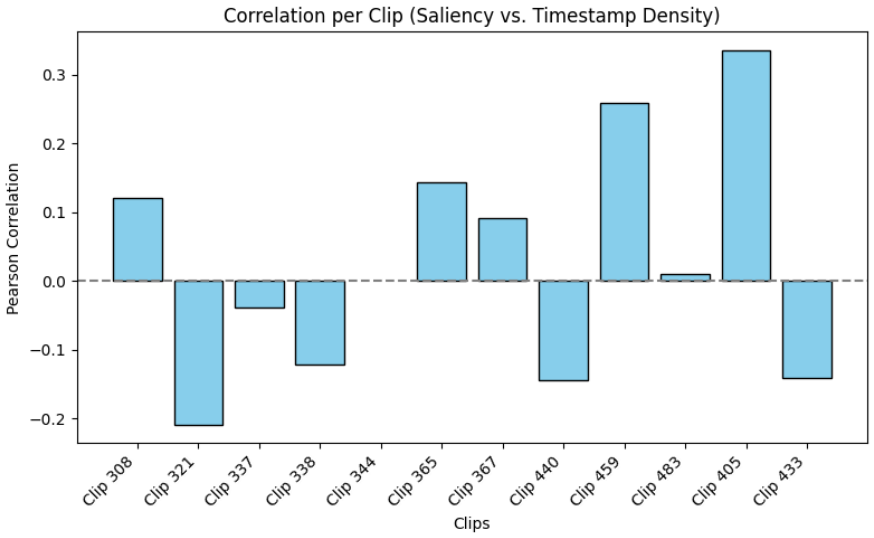


Figure 5.15: Pearson Correlation between saliency and timestamp density

We investigated another case where the participant correctly identified clips as not-depressed. We compare the alignment of their influential timestamps with the saliency map generated from all the segments that are contained in the 8.5 seconds video clip that were correctly identified as non depressed from the model.In Figure 5.16 we can see an example of a clip that was correctly classified from participants as non-depressed and the saliency map of this clip containing only segments that were correctly classified as non depressed by the model.

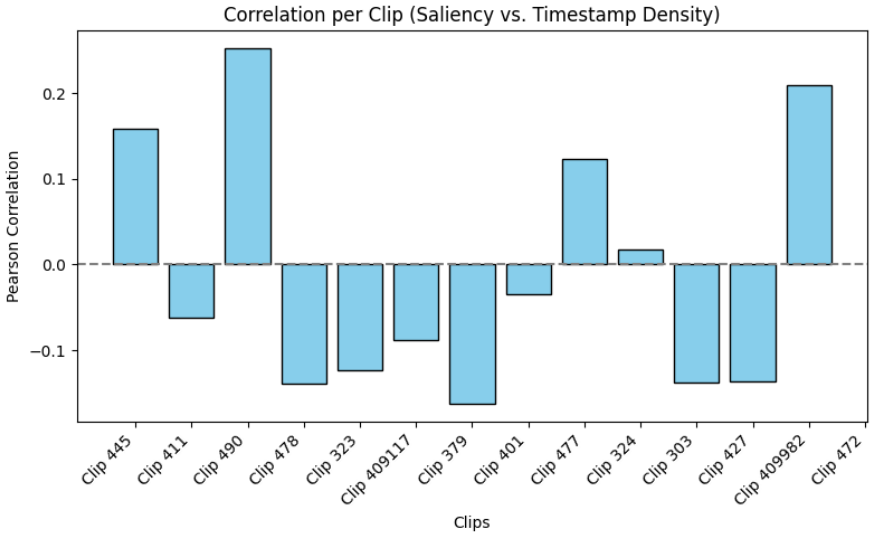


Figure 5.17: Perason Correlation between saliency and timestamp density

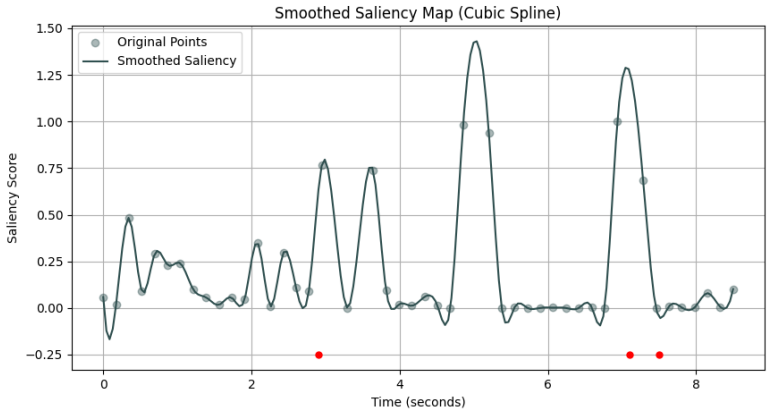


Figure 5.16: Selected timestamps along with saliency map for participant 445

Similarly as above, Figure 5.15 plots the Pearson correlation (y-axis) for each clip (x-axis) between the model's saliency distribution of TN segments and the density of participants' chosen timestamps when they correctly classified a participant as non-depressed. Overall, the correlations vary substantially across the clips. Several clips exhibit positive correlations (e.g., Clip 480, Clip 472, and Clip 40982), indicating that participants may have chosen timestamps for non-depression that tend to coincide with segments the model deemed salient when classifying these clips as non-depressed. Conversely, other clips show near-zero or even negative correlations (e.g., Clip 427, Clip 24, and Clip 478), suggesting that in those instances, participants may have selected timestamps that do not align—or may even diverge—from the segments to which the model assigned higher saliency. Again just like for the previous case, we use p-value as a measure of confidence in the statistical significance of the observed correlation. Similarly to the previous example, none of the above correlation has a p-value below the conventional significance threshold (0.05) indicating that the observed correlations could be due to random chance rather than a true underlying relationship.

The final step of our investigation is to determine whether the model's highlighted segment indirectly shapes participants' classification decisions. Figure 5.18 compares how many participants correctly identified depression across three different "cases," each representing a distinct position of the model's salient segment (e.g., at the beginning, middle, or end of the clip). The bar chart illustrates the proportion of participants who correctly identified the clips as depressed (true positive classifications) ac-

cording to the location of the salient segment. The three cases represent different positions of the model-identified salient segment: at the beginning (Case 1), at the end (Case 2), and in the middle (Case 3). From the figure, Case 1 shows the highest true positive rate, suggesting that when the model's salient segment appears early in the clip, participants are more likely to recognize signs of depression. Case 2 exhibits a moderately high rate, while Case 3 shows the lowest proportion of correct identifications.

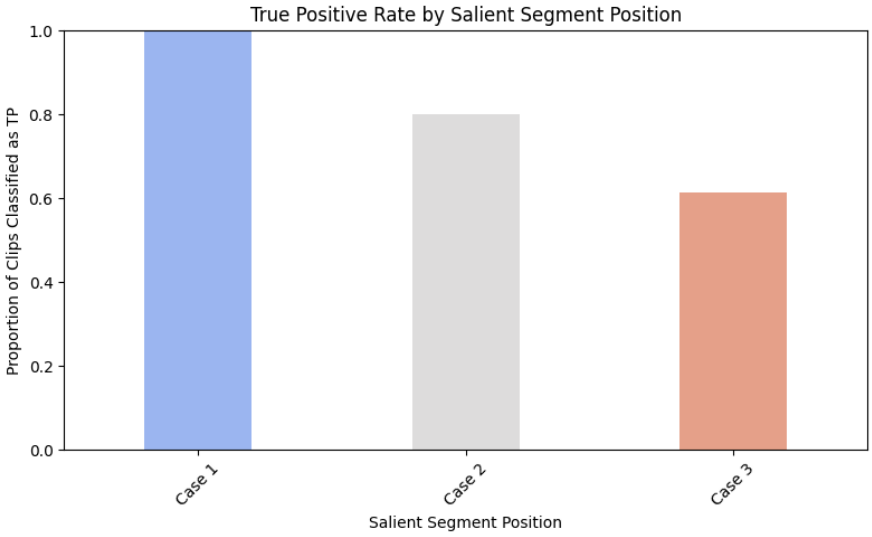


Figure 5.18: Classification rate per position

A *logistic mixed-effects model* was used to test the statistical significance of these results. The model included random intercepts for both Participant ID and Clip ID. The analysis included 34 observations from 17 participants and 10 clips. The model estimated a significantly higher probability of correct classification when the salient segment occurred in Case 1 compared to Case 2 (β = –2504.10, SE = 95.21, p < 0.001) and Case 3 (β = –2518.51, SE = 94.12, p < 0.001). The intercept (representing Case 1) was also significant (β = 2535.04, SE = 93.75, p < 0.001).

5.2.3. What facial/voice features do human identify in AI-selected salient moments and in their own selected salient moments?

This section of our research aims to compare the cues that participants receive in the model's selected moment to those which they receive in the moment that they find most persuasive themselves. By examining which facial cues and vocal cues are received in each case, we can determine whether the AI's chosen cues match human intuition or if participants pick up on different signals entirely. Additionally, we will investigate the features that the humans perceive and determine whether they are able at the first place detect meaningful facial or vocal features from short video clips.

Again for more detailed analysis, we divided the different clips in 4 categories. The ones that depict depressed individuals and were classified correctly (TP), the ones that depict depressed individuals and were classified incorrectly (FN), the ones that depict non-depressed individuals and were classified correctly (TN) and the ones that depict non-depressed individuals and were classified incorrectly (FP). This is done for both facial and voice features as they will be analyzed separately.

Facial Features analysis: Frequency/Co-occurance

The UpSet diagram below Figure 5.19 illustrates the facial features that participants selected in clips they correctly classified as depressed. On the left side, the horizontal bar chart ("Feature Count") shows how frequently each individual facial feature appeared across these clips, with higher bars indicating more frequent selections. On the top, the vertical bars ("Intersection Size") represent how many clips share particular combinations (intersections) of these features. For instance, a bar of height 2 indicates

that a specific combination of features was present in two clips, whereas a bar of height 1 indicates that a given combination was present in only one clip. Several facial expressions such as “Brow lowering,” “Outer edges of eyebrows raised,” “Repeated blinking,” and “Lips pressed together” appear relatively often, whereas “Smiling” or “Slight smirk” are selected less frequently. In many cases, only one or two facial features appear together, suggesting that participants often rely on a small set of facial cues to reach a conclusion of “depressed”.

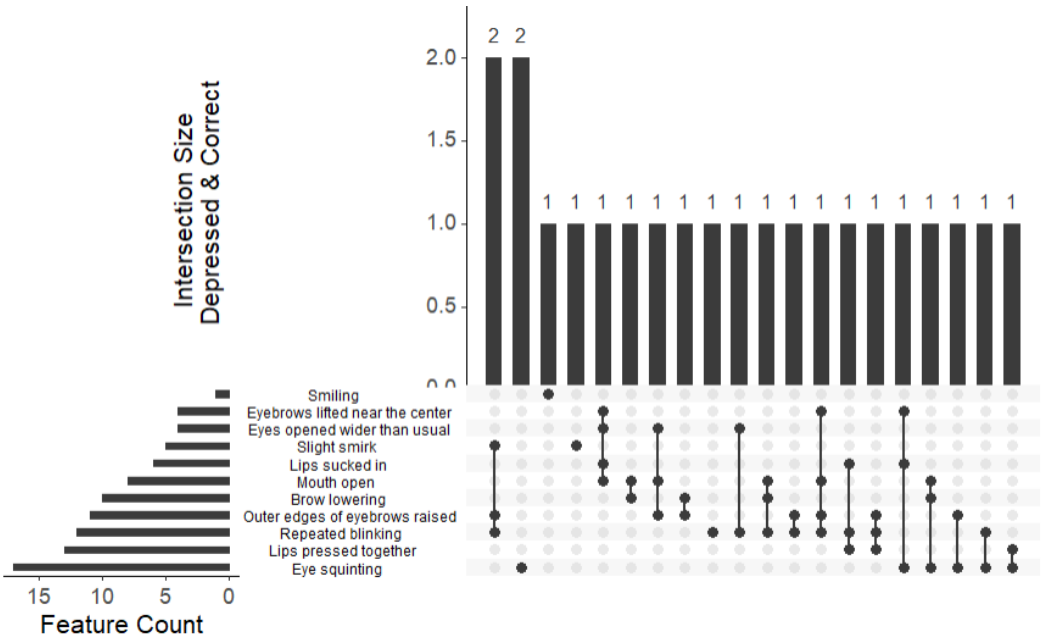


Figure 5.19: UpSet diagram for Clips that were correctly classified as depressed

In Figure 5.20 we illustrate with another UpSet diagram the facial features that were more influential for participants to classify a person as non-depressed even though they were. In this subset, participants relatively often noted cues such as “Repeated blinking,” “Eye squinting,” and “Brow lowering,” whereas features like “Eyebrows lifted near the center” or “Lips sucked in” appear less frequently (shorter bars at the top).

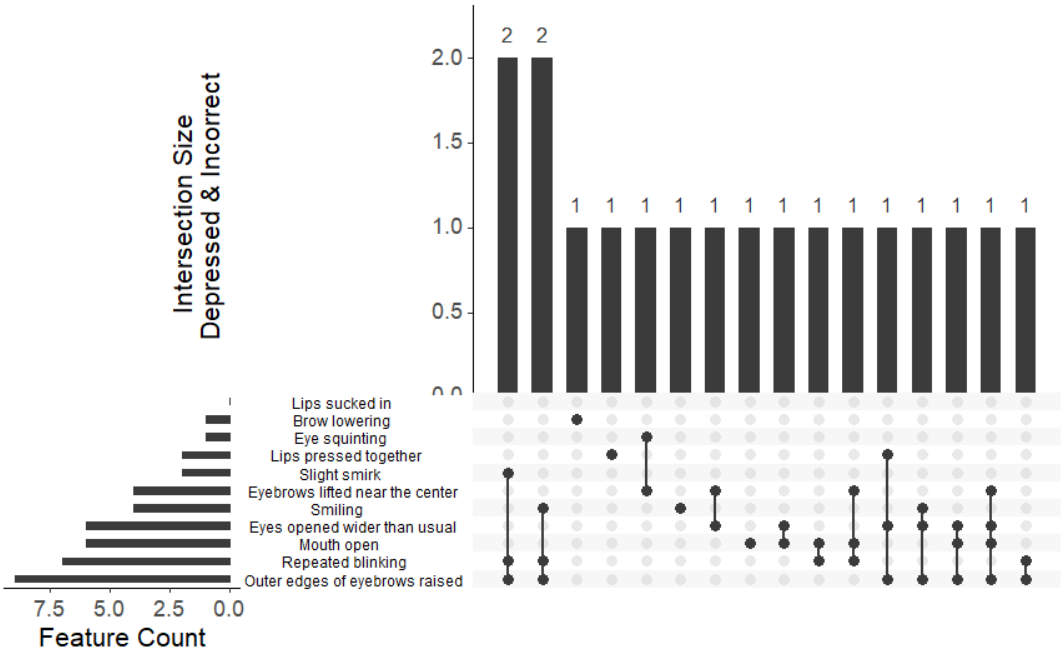


Figure 5.20: UpSet diagram for Clips that were incorrectly classified as depressed

Now let's investigate the cases where the clips represented a non-depressed individual. Figure 5.21 shows the facial features that were cues for the participant to choose non-depression. In this subset, "Repeated blinking" and "Eyebrows lifted near the center" appear most often, followed by features such as "Mouth open" and "Smiling." Meanwhile, features like "Brow lowering" or "Lips sucked in" are less frequent. Again, most intersections involve only one or two features, suggesting that participants typically relied on just a few cues to recognize a non-depressed presentation.

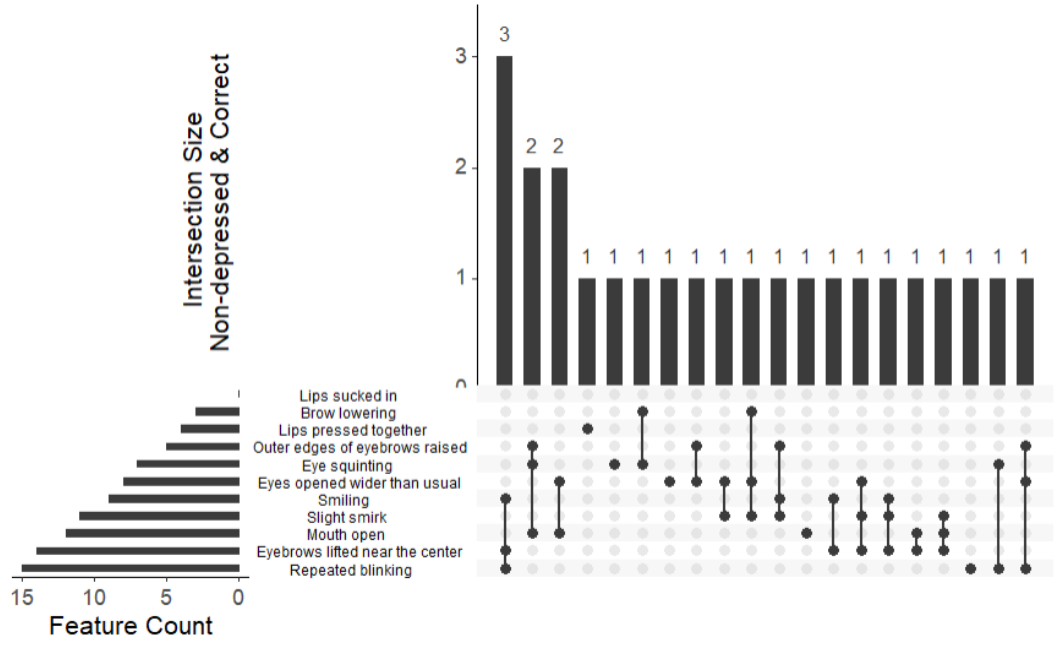


Figure 5.21: UpSet diagram for Clips that were correctly classified as non-depressed

In Figure 5.22 the UpSet diagram shows the most frequent cues and combinations for cases where

the participants classified the clip as non-depressed but they were incorrect. ”Repeated blinking” was chosen more frequently—followed by “Eye squinting,” “Brow lowering,” and “Lips pressed together.” Other features like “Mouth open,” “Slight smirk,” and “Eyebrows lifted near the center” occur less often (they have shorter bars).

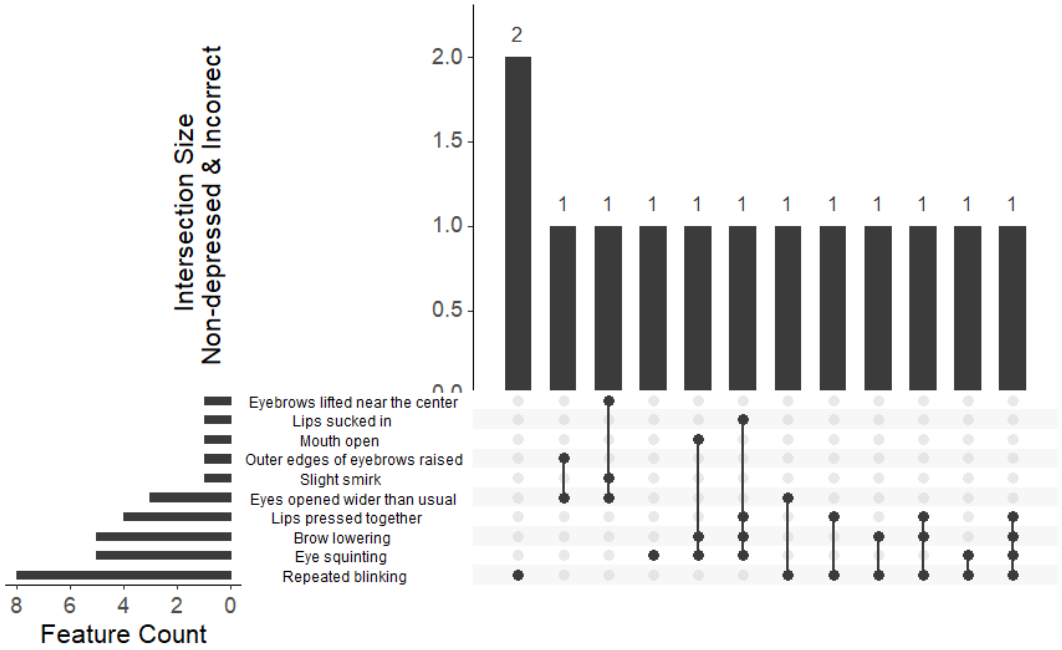


Figure 5.22: UpSet diagram for Clips that were incorrectly classified as depressed

Vocal Features analysis: Frequency/Co-occurance

The UpSet diagram below Figure 5.23 illustrates the vocal features that participants selected in clips they correctly classified as depressed. In this subset, features such as “Slow speech with long pauses,” “Monotone or flat voice,” “Speech that sounds hesitant or unsure,” and “Reduced Loudness” appear most often, while expressions like “Increased Speech Rate,” “Laughter or lighthearted tone,” and “Speech that is fluid and uninterrupted” show relatively low counts. Several columns have bars of height “4,” indicating that the same combination of features appeared in four different clips, suggesting a recurring cluster of signs (e.g., slow, hesitant, or monotone speech) that reliably led participants to identify those clips as depressed.

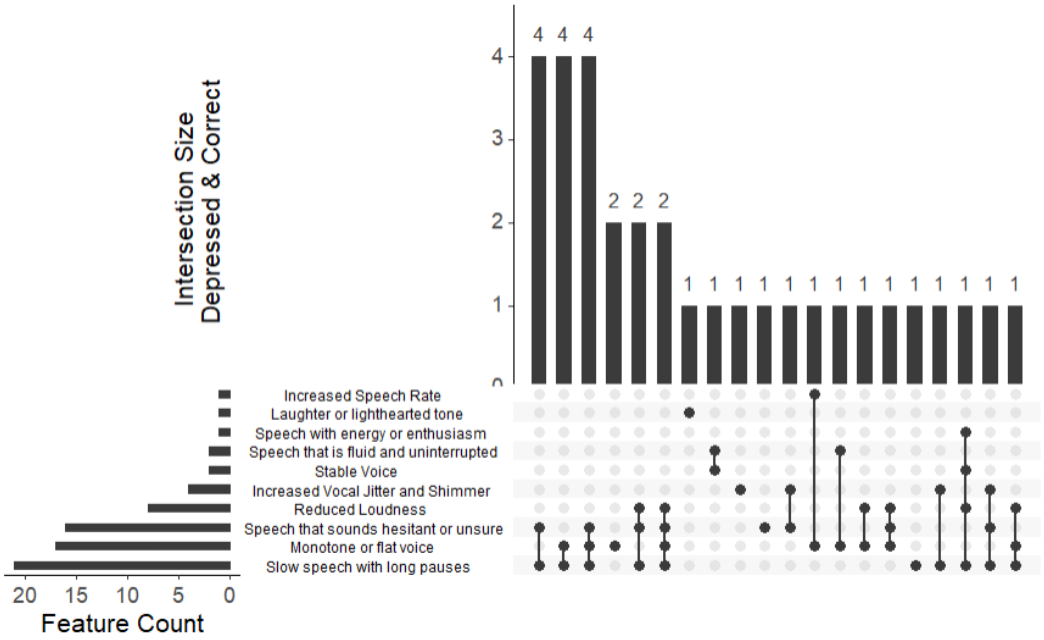


Figure 5.23: UpSet diagram for Clips that were correctly classified as depressed

This UpSet diagram Figure 5.24 displays the vocal features that participants noted in clips where the ground truth was depressed but participants incorrectly judged them as non-depressed—in other words, false negatives (FN). In this subset, participants often reported more “positive” or “functional” vocal qualities—such as “Speech that is fluid and uninterrupted,” “Speech with energy or enthusiasm,” and “Stable Voice”—which presumably led them to believe the speaker was not depressed. Less commonly cited were cues traditionally associated with depression, including “Reduced Loudness” and “Slow speech with long pauses.

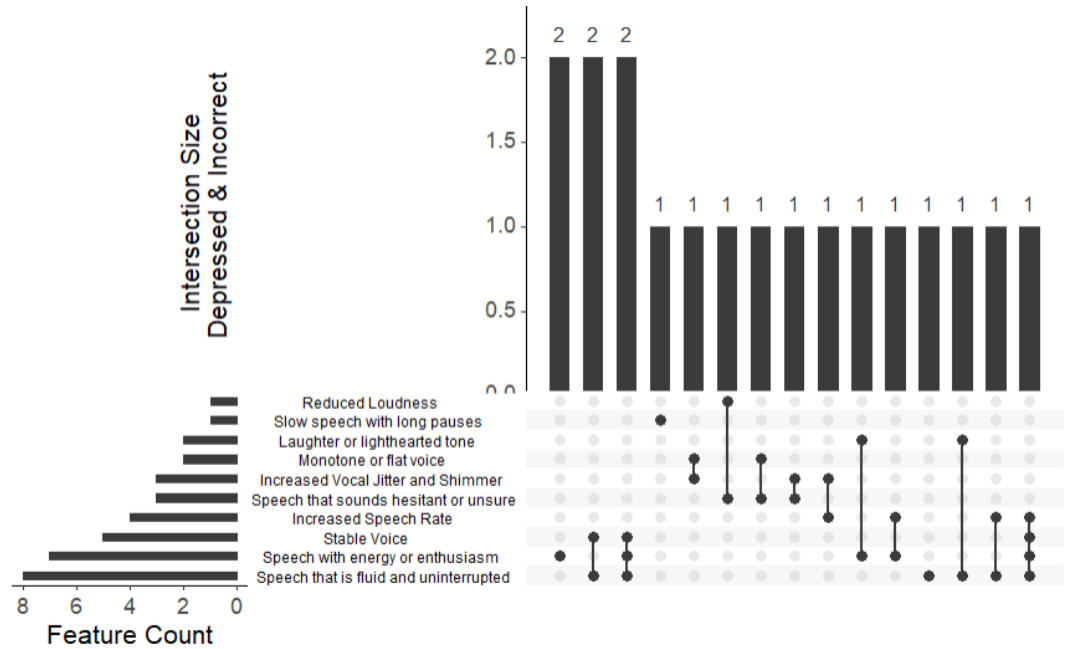


Figure 5.24: UpSet diagram for Clips that were incorrectly classified as non-depressed

This UpSet diagram **??** shows which vocal features participants identified in non-depressed clips that

they correctly labeled as non-depressed (true negatives). In this subset, participants commonly noted “Stable Voice,” “Speech that is fluid and uninterrupted,” and “Speech with energy or enthusiasm,” suggesting they interpreted these more upbeat or steady vocal qualities as signs that the speaker was not depressed. Features such as “Monotone or flat voice,” “Slow speech with long pauses,” and “Reduced Loudness” appear less frequently (nearer the top, with shorter bars), indicating that the cues participants often associate with a depressed speaking style were seldom observed in these clips.

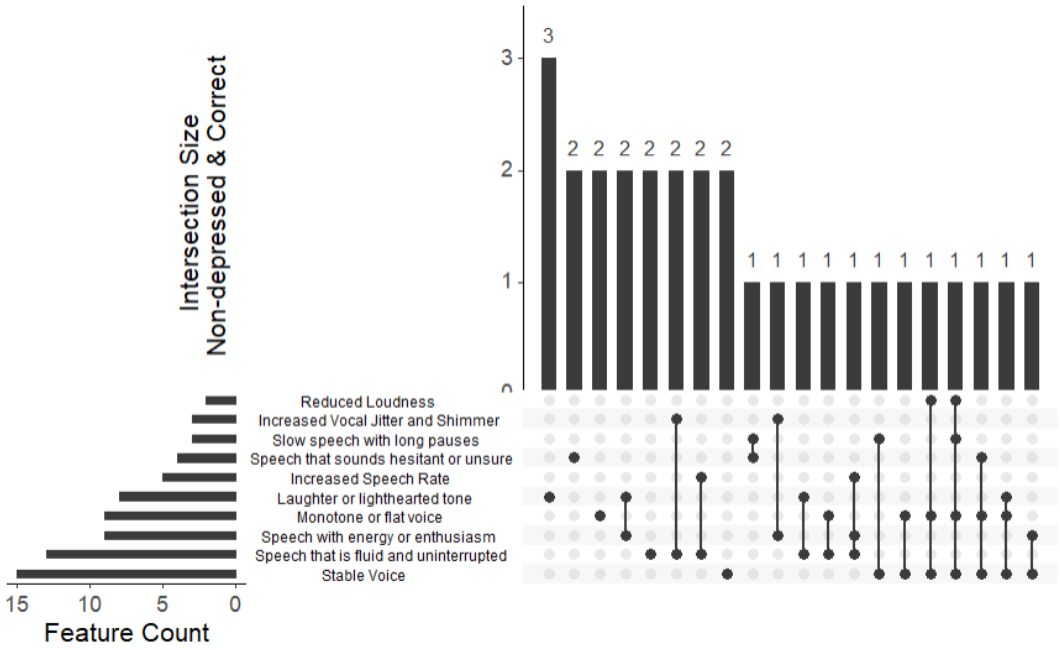


Figure 5.25: UpSet diagram for Clips that were correctly classified as non-depressed

Figure 5.26 depicts the vocal features participants noted in clips that were actually non-depressed yet incorrectly labeled as depressed—i.e., false positives. In this subset, participants often highlighted cues like “Speech that sounds hesitant or unsure,” “Reduced Loudness,” and “Monotone or flat voice,” which they may have interpreted as indicative of depression. Features such as “Laughter or lighthearted tone” or “Increased Speech Rate” appear less frequently. More frequently, hesitant speech and vocal jitter and shimmer were chosen together.

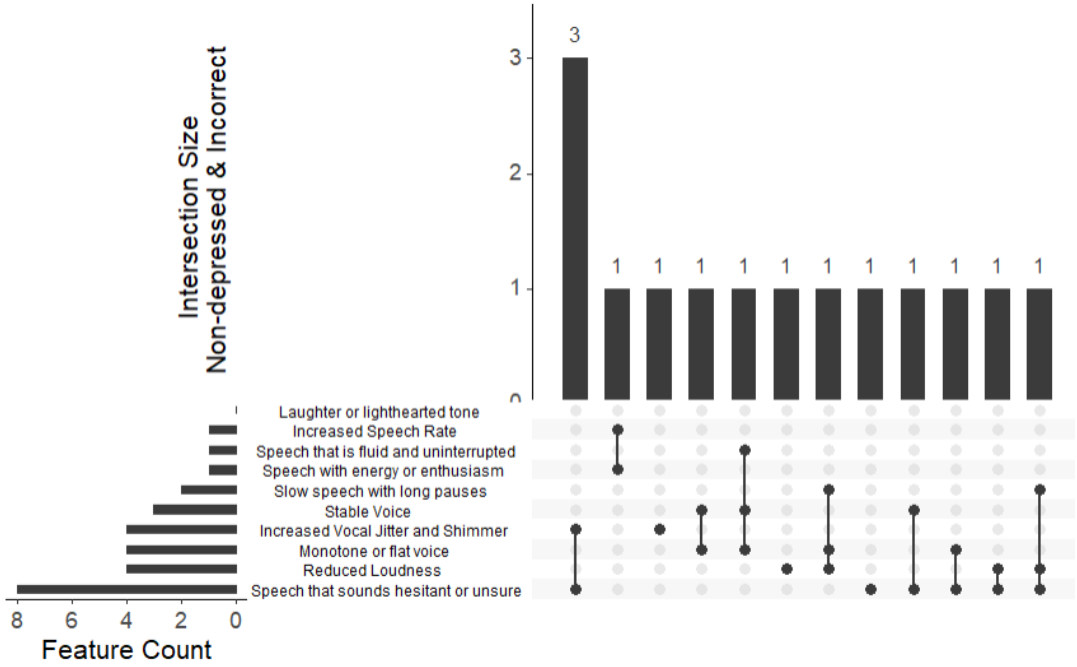


Figure 5.26: UpSet diagram for Clips that were incorrectly classified as depressed

Facial Features analysis: Selected moment vs Salient moment

In order to visualize the alignment between the facial features that participants selected versus the features the model deemed salient, we generated spider (radar) charts. Each radial axis corresponds to a different facial feature (e.g., “Brow lowering,” “Smiling”), and the distance from the center reflects how strongly or frequently that feature appears. The yellow polygons represent the features observed in the participant’s selected salient moment and the red polygon represents the features observed in the model’s selected salient moment. By examining the overlap or divergence of the two polygons, we can see whether the AI-highlighted segment emphasizes the same facial cues that participants naturally identify, shedding light on how well the model’s reasoning aligns with human perceptions. For the purpose of this thesis we will analyze these four scenarios:

- *TP-TP (Participants Correctly Classified Depression, Model Salient Segment Also Correctly Classified as Depressed)*

  The red (Selected) and yellow (Salient) polygons in Figure 5.27a show a relatively high overlap in certain features. For example, both participants and the model commonly highlighted expressions such as [e.g., “brow lowering,” “repeated blinking”] as important. The overall similarity in the shapes and magnitudes suggests that, in these clips, human and model judgments converged on similar facial cues for depression.

- *TP-FN (Participants Correctly Classified Depression, Model Salient Segment Incorrectly Classified as Non-depressed)*

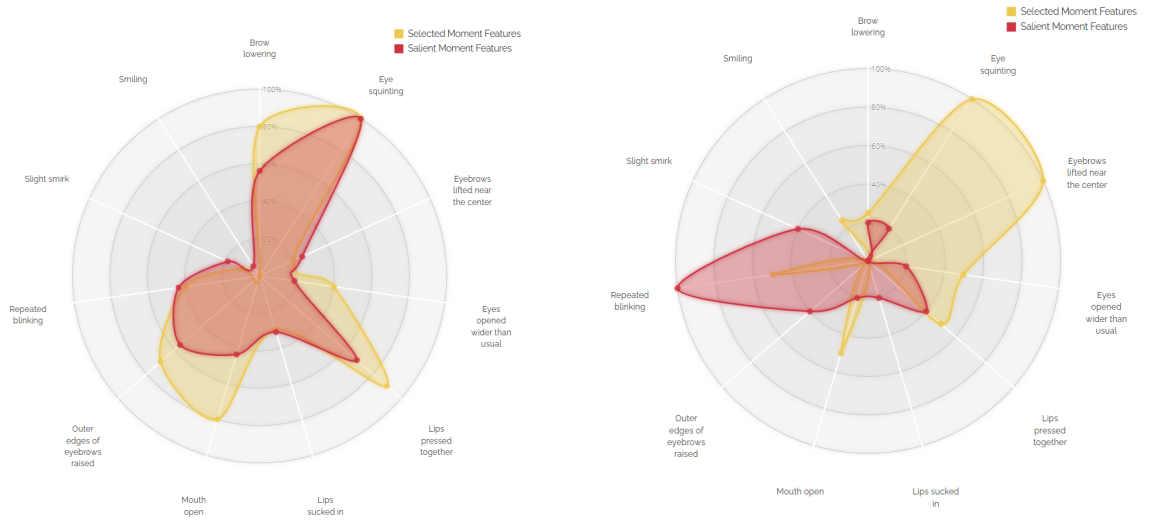
  Here in Figure 5.27b, the red polygon (participants’ selected features) appears to emphasize cues that the model did not capture as strongly in its yellow polygon. While participants noted [e.g., “eye squinting,” “lips pressed together”] with relatively higher intensity, the model’s salient features for these clips remained lower on those same axes.

- *TN-TN (Participants Correctly Classified Non-depression, Model Salient Segment Also Correctly Classified as Non-depressed)*

  In Figure 5.28a, the shapes for Selected and Salient often show moderate overlap in features generally associated with neutral or positive affect, such as [“smiling” or “outer edges of eyebrows raised”]. The model’s salient cues and the participants’ chosen features both indicate a lack of strongly “depressive” expressions, leading to accurate non-depressed classifications.

- *TN-FP (Participants Correctly Classified Non-depression, Model Salient Segment Incorrectly Classified as Depressed)*

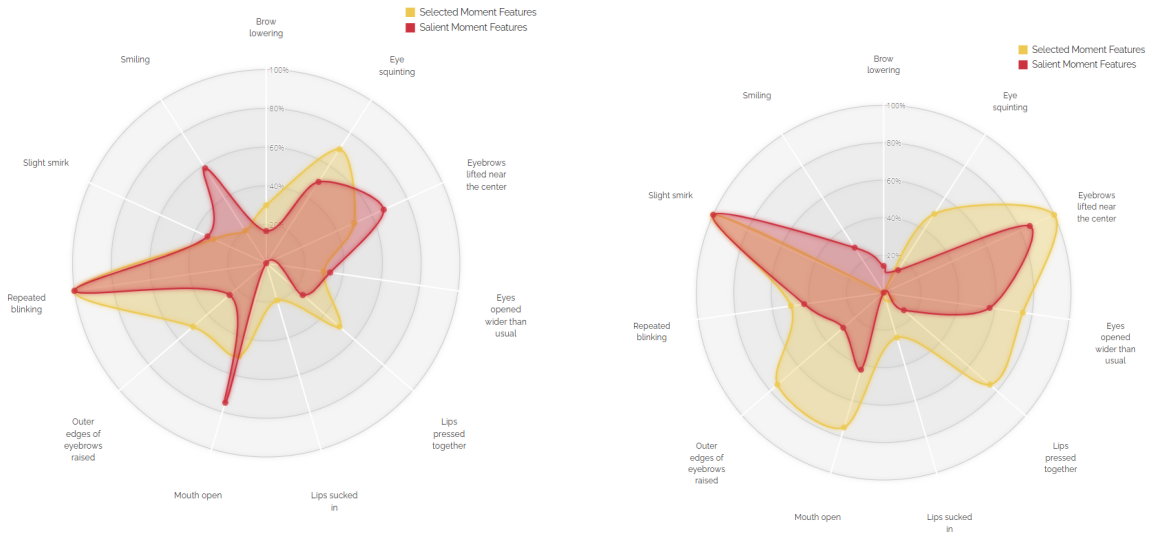
  In Figure 5.28b the red polygon typically shows lower values for tension-oriented features, reflecting the participants' judgment that the clips are non-depressed. However, the model's yellow polygon includes relatively higher values on certain cues [e.g., "brow lowering," "eye squinting"], suggesting it may have over-interpreted mild or ambiguous expressions as depressive signals.



(a) Salient segment identified correctly as depressed by the model

(b) Salient segment identified incorrectly as non-depressed by model

Figure 5.27: Spider diagrams for clips that were correctly identified as depressed from participants



(a) Salient segment identified correctly as non-depressed by the model

(b) Salient segment identified incorrectly as depressed by model

Figure 5.28: Spider diagrams for clips that were correctly identified as non-depressed from participants

Vocal Features analysis: Selected moment vs Salient moment

Similarly we investigated the differences for the vocal features again with the use of a spider diagram. Each radial axis corresponds to a different vocal feature (e.g., "Monotone or flat tone," "Stable voice"), and the distance from the center reflects how strongly or frequently that feature appears. The blue polygons represent the features observed in the participant's selected salient moment and the green

polygon represents the features observed in the model's selected salient moment. Again we will analyze these four scenarios:

- *TP-TP (Participants Correctly Classified Depression, Model Salient Segment Also Correctly Classified as Depressed)*

  Both polygons (Selected vs. Salient) in Figure 5.29a show high overlap in features typically associated with depression, such as ["Slow speech with long pauses," "Monotone or flat voice," and "Reduced Loudness"]. The similarity of the polygons indicates that in these clips, both participants and the model converged on vocal cues commonly linked to depressive affect.

- *TP-FN (Participants Correctly Classified Depression, Model Salient Segment Incorrectly Classified as Non-depressed)*

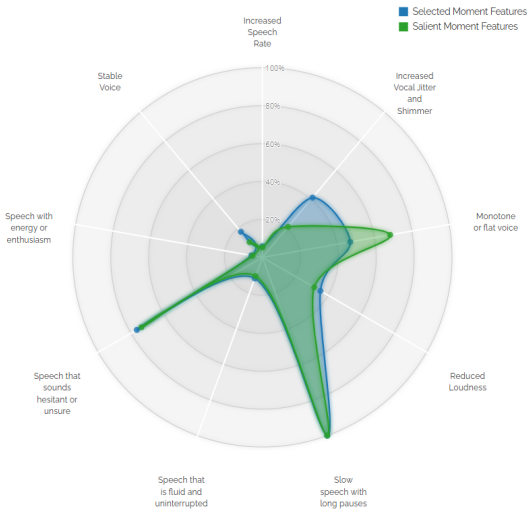
  Here in Figure 5.29b, the participant-selected polygon (e.g., the blue shape) tends to emphasize certain depression-related cues that the model's salient polygon (e.g., green) does not highlight. For instance, participants may frequently note ["Increased vocal jitter and shimmer"], whereas the model's salient moment values are comparatively lower on that axis. There is no big overlap between the participant's selected moment features and the features observed in the model's selected salient moment.

- *TN-TN (Participants Correctly Classified Non-depression, Model Salient Segment Also Correctly Classified as Non-depressed)*

  In this Figure 5.30a, both polygons generally show lower values for depression-linked cues (like slow, flat, or hesitant speech) and relatively higher values for more neutral or positive-sounding qualities (e.g., ["Speech with energy or enthusiasm," "Fluid and uninterrupted speech"]). This indicates a consensus between participants and the model that no strong vocal indicators of depression were present.

- *TN-FP (Participants Correctly Classified Non-depression, Model Salient Segment Incorrectly Classified as Depressed)*

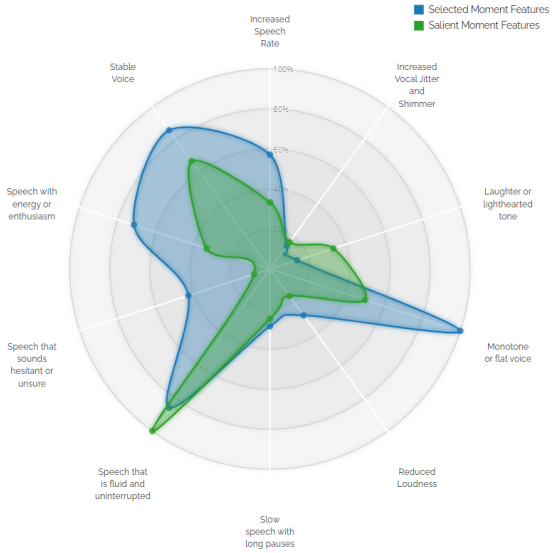
  Again in Figure 5.30b we observe a mismatch in the two polygons as expected. In the participant-selected segment, features such as "Speech with energy or enthusiasm" and "Stable voice" have higher relative values, whereas in the model's salient segment, features like "Speech that sounds hesitant or unsure" and "Slow speech with long pauses" show moderately elevated values. Overall, the participant's polygon covers fewer depression-associated features than the model's polygon, which emphasizes some cues commonly linked to depressed speech but did not result in a correct final classification.



(a) Salient segment identified correctly as depressed by the model

(b) Salient segment identified incorrectly as non-depressed by model

Figure 5.29: Spider diagrams for clips that were correctly identified as depressed from participants



(a) Salient segment identified correctly as non-depressed by the model

(b) Salient segment identified incorrectly as depressed by model

Figure 5.30: Spider diagrams for clips that were correctly identified as non-depressed from participants

6

Discussion

6.1. Results interpretation

Our study examined how human judgments and model-computed saliency maps align when identifying depression cues in short clips. Overall, our results reveal a complex picture. As clearly stated in the Results chapter, the classification results show that approximately two thirds (66.7%)of the clips were correctly labeled as either depressed or non-depressed, while one third (33.3%) were misclassified. Although the overall accuracy suggests that the model or approach captures some relevant features of depression, the one-third error rate indicates considerable room for improvement. We will delve a bit deeper in the analysis and interpretation of the results.

Inter rater reliability

The fair AC_1 (≈ 0.40) does not reflect an inability to detect depression per se-each rater on average classified clips above chance-but rather a lack of consensus about which specific clips carry clear depressive signals. In other words, raters tended to disagree on the ambiguous cases rather than all missing the same clips. That is because, some clips contained only moderate signs of depression that potentially contain more subtle cues and is more probable to create discrepancy between the participants. Additionally, this dispersion likely stems from the varied ways individuals weigh verbal versus nonverbal cues, idiosyncratic thresholds for “depressed” affect, and the absence of a shared decision rubric. Cognitive load may also have played a role since the same clips for some participants may appeared at the beginning of the experiment while for other it may appeared at the end, where the participants are more prone to mistakes since they have already performed a lot of cognitive and mentally straining tasks.

Confidence analysis

Figure 5.3 illustrates the distributions of confidence ratings for depressed and non-depressed clips. Observing the two distributions, the median confidence rating for depressed clips appears higher than that for non-depressed clips, suggesting participants generally provided higher confidence scores when a clip actually depicts depression. However, the overlapping ranges in both distributions indicate that this confidence is not absolute: some non-depressed clips still receive moderately high confidence scores, and some depressed clips receive lower confidence ratings. Thus, while a higher median confidence for depressed clips implies that participants generally feel more certain about identifying depression when it is present, the overlap shows there is still considerable ambiguity in how individuals assess these cues. This ambiguity highlights the inherent challenges in identifying depression based solely on observable cues and suggests that while confidence ratings provide useful information, they may need to be supplemented with additional diagnostic methods to enhance accuracy.

A subsequent linear mixed-effects analysis confirmed these observations.The linear mixed-effects analysis indicates that clips labeled as “depressed” received higher confidence ratings (on average 1.78

points more on the scale of 1 to 10) than those labeled “non-depressed.” This suggests that participants not only perceived stronger cues when judging a clip as depressed, but they also felt more certain in that judgment. One possibility is that depressive signals—whether facial, vocal, or contextual—stand out more distinctly to observers, resulting in greater confidence. Another explanation is that participants may be more cautious about missing a potential sign of depression, leading them to assign higher confidence once they suspect depression.

Interestingly, the random effect for participant was effectively zero, implying that individual differences in how confident participants felt did not vary substantially. In other words, participants showed a fairly consistent pattern in how they rated clips. The main variability was at the clip level, with some clips generally evoking higher or lower confidence across the board. This finding highlights the role of the clip’s inherent characteristics—such as the clarity of depressive indicators—rather than major differences in individual participants’ tendencies to give higher or lower confidence ratings. It also highlights the variability in depression manifestation among different individuals, making some of them more difficult to ’read’ than others as they have a smaller or different expression palette. Overall, the implication of these results is that people do sense something in depressed clips that leads them to higher confidence overall. It’s not purely random guessing.

The results in Figure 5.4 combined with the linear mixed-effects analysis, suggest that participants exhibit greater certainty when they accurately identify a clip as depressed, implying some degree of insight into their own accuracy. At the same time, the presence of lower confidence for false positives highlights that misclassification is associated with reduced certainty. This suggests that participants may have some metacognitive insight into their own decision accuracy—they ”feel” more certain when their judgment is right.Taken together, these findings underscore the importance of examining not only the accuracy of classification judgments but also the confidence with which those judgments are made, as it can offer additional clues about how observers process cues related to depressive symptoms.

Alignment with salient moments

Before delving into analyzing the salient moments, it is important to recognize that the model’s salient segments do not necessarily correspond to clear or strong indicators of depression. Instead, the model identifies salient moments based on local entropy changes, indicating segments that introduce uncertainty or surprise to the model’s predictions. This means a salient segment may represent various phenomena, such as sudden changes in behavior, shifts in facial expression, vocal characteristics, or other contextual factors, and not exclusively explicit signs of depression.

This is something that we took into consideration in the visualizations in Figure 5.5, Figure 5.6, Figure 5.7 and Figure 5.8. When a moment is salient, we need also to take into account whether or not the segment it belongs to was correctly or incorrectly classified by the model. This distinction gives us valuable information.

- **True Positive (TP)**: The segment is actually from a depressed participant, and the model correctly identifies it as depressed. Interpretation: The model’s salient moment might have captured genuinely depressive features, making it easier for humans to align.
- **True Negative (TN)**: The segment is actually from a non-depressed participant, and the model correctly identifies it as non-depressed. Interpretation: The salient moment might represent clear signals of a healthy emotional state.
- **False Positive (FP)**: The segment is non-depressed, but the model incorrectly classified it as depressed. Interpretation: If the model’s salient moment contains any identifiable features, they may reflect ambiguous or misleading behaviors that led the model to mistakenly infer depressive characteristics.
- **False Negative (FN)**: The segment is from a depressed participant, but the model incorrectly classified it as non-depressed. Interpretation: Here, the salient moment likely signals unexpected patterns or uncertainty—possibly subtle, atypical depressive signs that were hard for the model to correctly interpret.

Interpretation of Figure 5.5

The observed patterns (Figure 5.5) highlight the complexity of interpreting salient segments identified

by the model. A slightly higher alignment between participants and the model for segments correctly classified (TP) suggests that when the model potentially identifies segments indicative of depression, these segments likely contain more recognizable or universal cues of depression that human observers also readily detect. However, even within these correctly classified segments, there remains noticeable variability in participants' chosen timestamps. Additionally, this slightly higher alignment can be coincidental.

In contrast, segments misclassified by the model as non-depressed (FN) exhibit slightly higher dispersion in participants' responses, reflecting higher ambiguity. This result is intuitive: since participants had already correctly classified these clips as depressed, they specifically sought cues associated with depression. The model, however, identified salient segments in these FN cases likely due to features inconsistent with depression, such as expressions or behaviors that appeared unexpectedly non-depressed to the model. Consequently, human participants, possibly focusing on depressive indicators, responded with timestamps more widely dispersed and misaligned from the model's FN segments.

Interpretation of Figure 5.6

The results presented in Figure 5.6 reveal important insights regarding the alignment of human and model-identified salient moments for non-depressed clips. In some clips, these circles are centered within the densest region of participant selections, suggesting that in those cases where the model may captures features—perhaps clear signals of a healthy state—the participants also capture features they consider non-depressive, suggesting the model's salient moment aligns with participant judgments. However, in other clips, the green circle sits on the edges or even outside the main distribution, indicating that the model's salient point—while ultimately leading to a correct classification—does not necessarily coincide with the majority of participant-selected timestamps.

By contrast, the red triangles (FP) represent cases where the model incorrectly flagged a non-depressed clip as depressed. These points sometimes lie well outside the violin's densest region, implying that the model focused on cues that participants generally did not find indicative of depression. Perhaps in these clips the cues that the model picked up (and misclassified as depression) are ambiguous or even misleading to human observers. This could indicate that the model is sensitive to features that do not necessarily trigger the human perception of depression. In other clips, however, the red triangle is not so far removed from participant selections, suggesting that the model and participants might have picked up on similar signals yet interpreted them differently since participants have correctly classified the clip as non-depressed while the model incorrectly classified the segment as depressed.

Overall, the figure shows that correct classification (TN) does not guarantee a better overlap between the model's salient point and the timestamps participants that are most informative, and incorrect classification (FP) does not always mean a stark deviation from participant perceptions. Instead, the degree of overlap varies clip by clip, highlighting that model saliency (based on entropy changes) and human-selected cues (based on subjective assessments) sometimes coincide and sometimes diverge.

Interpretation of Figure 5.7

The results shown in Figure 5.7 provide important insights into the complexity of identifying depression from short clips. Participants incorrectly labeled these individuals as depressed, indicating they perceived cues that resembled depressive behavior. Interestingly, some model-identified salient segments were also false positives, meaning the model similarly perceived signals incorrectly suggestive of depression. This shared missclasssification may reflect genuinely ambiguous or unclear behavioral signals within these clips, potentially causing confusion for both the model and human observers.

As we can see in the plot there is not clear distinction on whether one of the two categories TP and FN salient moment is closer or further to the participants' selected timestamps.Taken together, these patterns illustrate that correct classification (TN) does not guarantee strong alignment with human-selected timestamps, nor does a misclassification (FP) always mean the model's salient moment is far removed from human perceptions. Instead, the figure underscores how model saliency—derived from entropy changes—can sometimes coincide with human judgments, yet at other times emphasize moments that participants do not view as decisive.

Interpretation of Figure 5.8

The variability observed in Figure 5.8 aligns well with the nature of the data: these clips were misclassified by participants as non-depressed, indicating that participants may have struggled to identify clear depressive signals. The fact that some of the model's salient segments were also false negatives (FN)—segments the model did not identify as depressed—likely added to the confusion, as these segments may contain subtle or ambiguous behavioral cues not clearly indicative of depression.

Where the model correctly identified segments as depressed (TP), participant failed to recognize these indicators as clearly depressive. This pattern underscores the complexity of human judgments in subtle cases: even when clear indicators exist (as potentially recognized by the model), lay observers may overlook or misinterpret these signals, leading to classification errors.

Overall, in the diagrams above we see a clear discrepancy and not alignment between participants influential moments and the model's salient moment. It seems like in all the cases participants have their own personal reasons to believe something is salient that differ from each other and from the model most of the times. Later on we try to delve deeper in the alignment of salient moments with participants influential moments.

Error plots

TP-TP vs TP-FN

To further analyze and investigate the alignment of humans salient moments with the model's we use the raw error. The findings from the linear mixed effects model indicate that among clips where participants correctly classified depression, the temporal precision of their responses does not significantly differ between clips that the model also classified the salient segment as depressed (TP) and those that it did not (FN). Although the descriptive statistics suggest a modest shift in the average error (with TP clips showing slightly less negative error), the lack of statistical significance implies that participants' ability to accurately pinpoint the salient interval is relatively consistent, regardless of the model's classification of the top salient segment of the clip.

TN-TN vs TN-FP

These findings indicate that for non-depressed clips, where participants accurately identified the clip as non-depressed, the temporal precision of their responses—as measured by raw error—does not differ between clips that the model correctly classified their salient segments as non-depressed (TN) and those that were misclassified (FP). The significant negative intercept suggests an overall tendency toward a negative raw error, but the non-significant difference between TN and FP groups implies that, in terms of timing accuracy, the model's classification of the salient segment does not differentially impact participant performance.

Pearson Correlation

One salient moment itself may not give us enough information about the overall saliency of the clip. That is why we deepen the analysis further employing saliency maps. The Pearson correlation analysis shows us whether there is any correlation between the saliency values (taken from the saliency maps) and the timestamp density of the participant's responses. For this thesis we focused on cases where the participants correctly classified the clip (TP and TN). That is because we want to focus on whether the participants have strong influential cues to identify depression or non-depression and if the moments that influence the model do also influence the participants. We calculated the Pearson correlation diagrams using saliency maps of the clips. There are two types of saliency maps employed for this task. First saliency maps for depressed clips that only contain saliencies of segments that were correctly classified as depressed by the model (TP) and second saliency maps for non-depressed clips that only contain saliencies of segments that were correctly identified as non-depressed by the model.

TP-TP

The correlation analysis between participant-selected timestamps and model-generated saliency maps offers insight into how closely human observers' perceptions of depression cues align with those of the computational model. The presence of positive correlations in some clips suggests that, at least in

certain instances, participants and the model focus on similar visual or auditory cues that may be indicative of depression. These findings lend preliminary support to the idea that the model's saliency maps, which highlight regions it considers important for classification, capture cues that are also perceived as relevant by human annotators.

The variation across clips—ranging from positive to near-zero or negative correlations—underscores the complexity of human perception and the nuances of the model's learned features. It is possible that, for some clips, the model attends to subtle cues that participants do not consciously register or do not consider critical. Conversely, participants may pick up on broader contextual or narrative elements that the model does not capture, leading to lower or negative correlations. However, it is important to note that none of the correlations reached statistical significance ($p > 0.05$ for all clips), indicating that the observed associations could have occurred by chance. This lack of statistical power may, in part, be due to the limited number of participant observations per clip, which reduces the sensitivity of the correlation tests.

TN-TN

The correlation analysis between participant-selected timestamps and model-generated saliency maps provides insight into how closely human observers' perceptions of non-depression cues align with the model's learned features. In clips with positive correlations, participants and the model appear to converge on similar visual or auditory signals that affirm a non-depressed classification, offering preliminary evidence that the model's saliency maps capture aspects of the clips that humans also consider relevant. However, the presence of near-zero or negative correlations in other clips highlights the inherent complexity of interpreting non-depressive behavior. Participants may rely on broader contextual or social cues that the model overlooks, or the model may attend to subtle indicators that participants do not consciously register. These discrepancies emphasize the importance of refining model interpretability and exploring additional ways to bridge the gap between human judgment and machine-based saliency for more robust and transparent mental health assessments. Additionally, as mentioned before some clips may be more difficult for humans to classify than others since there may be the components of cultural bias or limited skills from the participants. However, just like in the previous case none of the correlations reached statistical significance ($p > 0.05$ for all clips), indicating that the observed associations could have occurred by chance.

Position bias

In Figure 5.18 there are illustrated the classification scores based on the salient segment positions. The scores are about True positive classifications meaning how many clips that are depressed were correctly classified as depressed based on the salient segment position. Since these clips' salient moment were correctly classified as depressed by the model (true positives), the salient segment in each case likely contained distinct depression cues. The higher true positive rate in Case 1 implies that when salient signals appear near the beginning, participants may find it easier to detect depressive indicators. Conversely, when the salient segment is in the middle (Case 3), participants might have more difficulty noticing or recalling those cues by the time they make their judgment. Interestingly, Case 2—where the salient segment is at the end—has a moderately high rate, which could be because participants are left with an impression of the depressive cues they see last, potentially influencing their final decision.

It is also possible that some participants even if they did not consciously select the model's salient moment as the most influential,they were nevertheless subconsciously influenced by it when making their final assessment. This discrepancy implies that the cues participants believe are most important may differ from those emphasized by the model. Participant may not highlight segments as important if these segments contain more subtle cues that even though they may subconsciously influence them they are not that evident to them.

To investigate the case that those results are due to random variation we use a *logistic mixed-effects model*. After including both participant and clip as random effects, the logistic mixed-effects model suggests that salient segment position strongly influences participants' ability to correctly identify depressed clips. However, the model yielded extremely large coefficients and standard errors, indicative of quasi-complete separation — a situation in which a predictor variable (in this case, Salient Position

Type) nearly perfectly separates the binary outcome. This is likely due to data sparsity since there are is a small number of observations. This leads to unstable estimates and suggests that while the trend is meaningful, the statistical model may be overfitting, and the results should be interpreted with caution. Additional data or more balanced designs may be necessary to produce stable, reliable effect estimates.

Feature Analysis

Facial features

The findings for TP clips align with prior literature, which reports that depressed individuals tend to show more brow-related actions (e.g., AU1, AU4) and increased blink rate (AU45). As seen in the diagram, participants frequently selected “Brow lowering” and “Repeated blinking,” possibly reflecting signs of sadness, internal stress, or tension. Furthermore, fewer expressions related to happiness or positive affect (e.g., smiling) were observed, which is consistent with research indicating that depressed individuals typically display lower frequency and intensity of AU12 (lip corner puller).

However, it is notable that features like “Outer edges of eyebrows raised,” sometimes linked to surprise or fear, also appeared relatively frequently. While the literature does not strongly link outer brow-raising (AU02) to depression, its consistent presence in the current data may reflect participants’ perception of tension or distress in the brow region more broadly. Similarly, partial smiles or smirks (AU14) were reported, though less frequently, which aligns with the notion that genuine smiling is uncommon in depression. Although a smile might typically be seen as a sign of positive affect, its presence dis not necessarily dissuade participants from concluding that a clip was depressed. Participants may have considered other salient cues, such as tension in the eyes or brow area, or voice cues to be more indicative of depression, thus “overriding” the presence of a smile. It is also possible that participants recognize that individuals with depression can still smile, whether as a coping strategy or a social mask. This finding aligns with broader discussions in depression research, indicating that facial expressions can be mixed or incongruent, and a momentary smile does not rule out an underlying negative mood.

In Figure 5.20, because these clips were truly depressed, yet participants did not recognize them as such, it suggests that some of the more ambiguous or even “neutral”-seeming cues (e.g., blinking, slight smirks, or widened eyes) may have obscured participants’ detection of depressive signals. Although “Brow lowering” or “Lips pressed together” can be signs of distress, these may not have been pronounced enough—or interpreted strongly enough—to override the potentially “normal” or less overtly sad expressions. This dynamic highlights how false negatives can arise when observers rely on highly stereotypical markers of depression (e.g., a flat or downcast face) and underestimate subtler facial indicators.

Compared to the depressed groups, the most frequently selected cues here in the TN case Figure 5.22—such as repeated blinking and eyebrows lifted near the center—may be interpreted by participants as neutral or relatively positive signs, thus reinforcing a “non-depressed” judgment. For instance, repeated blinking could be perceived as typical or comfortable behavior, rather than a sign of distress. Likewise, slightly lifted eyebrows or an occasional smile might signal a neutral or upbeat demeanor. In line with existing research indicating that depressed individuals often exhibit lower frequencies of genuine smiles, seeing a normal or slightly smiling expression may have led participants to conclude that these clips were not depressed. The relatively lower frequency of tension-oriented features (e.g., brow lowering) further supports that participants found fewer distress signals in these clips. Overall, these findings underscore that a handful of positive or neutral cues—especially around the eyes and mouth—were salient enough for participants to accurately identify these individuals as non-depressed. Features that scored high in TP observation such as Brow lowering and Lips pressed together now have low frequency scores which can indicate that are more likely to be linked as depressive cues from the participants.

Because the clips in Figure 5.22 were actually non-depressed, the presence of tension-oriented or ambiguous cues (e.g., frequent blinking, squinting, or lowered brows) seems to have prompted participants to over-interpret them as signs of depression. These findings highlight how relatively common facial expressions—such as blinking or a momentarily lowered brow—can be misconstrued as depressive indicators, especially if observers expect depression to manifest in any display of perceived stress or discomfort. Comparing these results to truly depressed clips (where features like prolonged sad

expressions are more prominent) could help clarify which signals genuinely reflect depressive affect and which simply convey mild tension or momentary discomfort.

The UpSet diagrams reveal that the co-occurring features selected by participants at a specific timestamp often do not match well-known, prototypical expressions. This likely reflects that participants are capturing fleeting, idiosyncratic moments rather than standardized expressions. In other words, the features chosen at a given moment may represent subtle or mixed cues that, while salient to the observer, do not neatly combine into the canonical expressions typically described in the literature. This variability underscores the complex, context-dependent nature of real-world affective signals, suggesting that what participants perceive as 'influential' may not always align with conventional expressions of depression

Vocal features

The findings in the TP case Figure 5.23 align with well-established notions about depressed speech patterns. Participants frequently highlighted slower, quieter, and more hesitant vocal qualities—characteristics widely reported in the literature as indicative of depressive states, such as monotone prosody, reduced loudness, and prolonged pauses. When such features co-occur (as shown by the intersection columns with higher frequencies), they may present a strong cumulative signal that the speaker is depressed. In contrast, features associated with more energetic or positive affect (e.g., laughter, higher speech rate, fluent speech) were seldom selected, implying that participants may regard these as counter-indicators to depression.

Overall, the results suggest that participants share a mental model of depressed speech marked by sluggishness, uncertainty, and diminished variation in pitch or volume. Recognizing these patterns can help researchers and clinicians refine automatic detection methods for depression and devise clearer guidelines for human observers—emphasizing the role that perceived "slowness," "flatness," and "hesitancy" play in identifying potential signs of depression from vocal cues alone.

Because these clips in Figure 5.24 were actually depressed but not recognized as such, it appears that relatively fluent, energetic, and stable speech overshadowed or masked more subtle indicators of depression. When participants heard someone speak with continuity, confidence, or even mild enthusiasm, they may have discounted any low-intensity signs of sadness or hesitation, thus labeling the clip as non-depressed. This pattern aligns with the idea that stereotypical "depressed speech" is slow, monotone, or hesitant; if a speaker deviates from that profile by sounding more fluid or confident, observers might miss underlying distress. Comparing these results to the Depressed & Correct cases highlights how strongly participants rely on slower, flatter speech as a depression cue. If those cues are absent—or if more positive-sounding cues are present—participants risk underestimating the possibility of depression.

Compared to depressed clips — where slower, flatter, and more hesitant voices were frequently identified — these non-depressed clips Figure 5.25 were characterized by vocal qualities that participants perceived as more lively, continuous, or confident. This distinction aligns with prior research suggesting that a smoother, more energetic delivery is generally associated with better emotional well-being. However, it is worth noting that some features typically linked to depression (like monotone speech) did surface in a handful of these non-depressed clips, highlighting the complexity of vocal expression and the possibility that certain cues can be context-dependent or interpreted differently depending on other co-occurring features. Overall, these results underscore that participants largely rely on markers of clarity, stability, and energy in the voice to make correct "non-depressed" judgments—a pattern consistent with lay intuitions about healthy or normal-sounding speech.

The findings in Figure 5.26 suggest that participants may perceive certain subdued or uncertain vocal qualities—such as hesitant, quiet, or monotone delivery—as red flags for depression, even when the speaker is not depressed. In other words, the absence of overt positivity or confidence might lead observers to mistakenly assume a depressed state. When contrasted with true negatives (correctly identified non-depressed clips), it becomes clear that features like stable voice and energetic speech are not as prevalent here, potentially fueling the misjudgment. This underscores the complexity of vocal-based depression detection: participants can over-attribute mild speech irregularities to depression.

Selected moment vs Salient moment

Facial features

These spider diagrams shed light on how both participants and the model identify key facial expressions associated with (or indicative of) depression. In the TP-TP and TN-TN scenarios, the convergence of features suggests that human observers and the model share similar heuristics—e.g., brow tension and blinking for depression, smiling or relaxed brow for non-depression. This agreement implies that the model can sometimes mimic human intuition effectively, lending preliminary support to its saliency mechanism. Of course, the features that participants observe in those salient moments does not necessarily equal what the model actually 'observed' in those salient moments. However, the matching of the features when both the model and the participants classified correctly (TN-TN and TP-TP) compared to the discrepancy of the features when the model and the participants classified differently (TN-FP and TP-FN) suggests that even with the bias, participants are able to differentiate between the cases where the salient moment is correctly or falsely classified by the model (hence possibly containing depression cues vs ambiguous cues).

Vocal features

These vocal-based radar charts complement the facial analyses, underscoring how participants and the model differ or agree on cues tied to depression. In the TP-TP and TN-TN cases, both human judgments and the model's salient segments align well, pointing to shared recognition of either "depressed" vocal markers (slower, flatter, and quieter speech) or "non-depressed" markers (clear, energetic, or uninterrupted speech). This alignment implies that the model can learn many of the same vocal signatures that participants intuitively associate with depression or its absence.

Similarly to before, the fact that in the cases of TP-FN there is discrepancy while in cases TP-TP there is alignment it show us that participants are able to pick up differences in voice features based on the meaning of the salient moment. Both facial and vocal diagrams reinforce the hypothesis that participants observations about the salient moment are valid and will help us shed light in what the model can use as cues to form a decision. The above results suggest that participants and the model seem to follow similar heuristics and value similar cues to form the right decision.

6.2. Implications

Our results have several important implications for improving the interpretability of AI-based medical diagnosis for depression through the identification and human evaluation of salient moments in video and audio data. First, the fact that humans correctly classify 66.7% of clips as depressed or non-depressed from short video segments (Sub-question 1) demonstrates that even brief clips can contain robust cues for depression.

Second, our findings reveal that human-identified salient moments and the model's salient moments diverge more often than they align (Sub-question 2). While there are instances of convergence—particularly in cases where both humans and the model correctly classify clips (TP-TP and TN-TN)—the predominant pattern is one of discrepancy. This divergence indicates that the model's method of deriving saliency (based on entropy changes) does not necessarily agree with what human observers may find salient. Even though such misalignment highlights a critical gap in interpretability, it does not directly suggest that the model does not follow relevant cues since there are several reasons to explain the discrepancy, discussed in the limitations section.

Third, in examining the specific facial and vocal features (Sub-questions 3a and 3b), our analyses show that when participants select influential moments, they tend to identify features that align with established literature—such as brow lowering in depressed faces and slower, more hesitant vocal patterns. In cases where human and model selections converge, the features are consistent with known depressive markers. However, discrepancies in cases like TP-FN and TN-FP indicate that even when participants are biased by their own interpretations, they still distinguish between influential moments when the model misclassifies the salient segment. This suggests that human evaluation is valuable not only for validating model outputs but also for highlighting the limitations of current explainability methods. Despite these limitations, this preliminary exploration into the relationship between human timestamp selection and model saliency highlights potential areas of convergence and divergence.

The discussed results point to the potential benefits of incorporating human expertise alongside algorithmic criteria. Rather than relying only on measures like entropy, the model could be refined by leveraging insights from clinicians or expert evaluators to pinpoint which parts of a video or audio clip truly reflect depressive symptoms. By recalibrating the model's outputs based on these human-identified cues, its predictions would be more closely aligned with clinical judgment, ultimately enhancing its accuracy and real-world utility.

6.3. Limitations

This study faces several noteworthy limitations that warrant careful consideration when interpreting the findings. First, all participants were lay individuals with no formal clinical training in mental health. Several studies (Slepian, Bogart, and Ambady, 2014; Rother et al., 2021) have shown that experts performance is not substantially higher that lay participants in identifying depression from visual and audio cues in thin-sliced scenarios like in our case (8.5 seconds clip). However, experts do tend to outperform laypeople in more nuanced or borderline cases, where depression is more mild and cues are more subtle, improving sensitivity Slepian, Bogart, and Ambady (2014). Second, the sample size was relatively small (17 participants), limiting the statistical power and the representativeness of the results for broader populations. For the Pearson correlation analysis the small number of participants per clip is an important limitation. With only three or four participants annotating each clip, individual differences in perception or annotation strategy can significantly influence the overall correlation values. Consequently, these findings cannot be generalized without caution.

Third, there was an imbalance in the experiments depicted clips, with fewer clips where the salient moment is part of a salient segment that is false-positive (FP) and false-negative (FN) than true-positive (TP) and true-negative (TN) cases. It is important to note that the number of clips with incorrectly classified salient moments (FN, FP) was about half that of correctly classified ones (TP, TN). This difference in sample size limits the statistical reliability of some observed trends might affect the robustness and generalization of the comparisons. Future studies could address this limitation by ensuring balanced sample sizes or employing weighting or normalization methods to better understand how alignment differs based on the correctness of the model's classification.

Another important limitation is that the facial expressions were presented through animation rather than real human faces, potentially making it a hard task for participants since subtle emotional cues are harder to interpret. An actual face of a real person would be much more easier and intuitive for participant to assess facial expressions on potentially giving us more valuable information. Additionally, regarding the video clips, there is a potential bias regarding the speech context. Even though the participants were prompted to not focus on the context of the speech and only in the voice and facial expression, such bias is difficult to overcome. So there is the possibility that the context of the speech subconsciously played a role in the participant's decisions.

Another concern arises from the sliding-window approach: as we discussed in the previous chapters a moment can be flagged as salient due to 2 points of interest. Either something important was removed in the 0.1 seconds of the slide or something important was added in the 0.1 that were added due to the slide. In this study we decided to focus on the second point of interest and around this we performed the analysis. But there may be salient points selected by the participants that are closer correlated to the other points of interest that were left out of the analysis. Furthermore, participants were asked to identify precise timestamps to a 0.1-second resolution, but the experimental interface allowed navigation only in one-second increments, introducing potential timing inaccuracies.

6.4. Future Improvement

Future research should address several key areas to enhance both the experimental design and the underlying model. First, expanding the participant pool—ideally including domain experts such as clinicians—would provide more robust and clinically informed assessments, potentially reducing subjective bias and increasing the reliability of human evaluations. Overall, the 66.7% accuracy indicates moderate performance. Future work could explore several avenues for improvement: (1) collecting additional data to increase the diversity of depressive presentations, (2) refining the model to focus on more discriminative features (e.g., combining facial, vocal, and contextual signals) and make it more

generalizable. Future research would benefit from increasing the number of participants per clip to reduce variability and provide a more robust test of alignment between human and model saliency. Additionally, designing more balanced experiments that contain the same number of FP and FN salient moments will be important for a more fair and generalizable comparison. These improvements will help bridge the gap between machine-driven analyses and human clinical intuition, ultimately enhancing the interpretability and practical utility of AI-based depression diagnostics.

7

Conclusion

In this work we set out to discover whether surfacing “salient moments” in audiovisual depression assessments-and then checking those moments against human judgments-could make AI explanations both more meaningful and more interpretable. We first showed that, although lay participants are only moderately accurate when forced to judge very short clips in isolation, the facial and vocal cues they cite (“eye-squinting,” “long pauses,” “monotone voice,” etc.) map directly onto well-established clinical markers. In other words, their selections are far from random—they echo the literature on depression indicators. We then compared those human-chosen features to the cues our model highlights at its own most salient time points. When the model correctly flagged a segment as “depressed,” there was a strong alignment between the participant-noted cues and the model’s information-theoretic saliency peaks. Conversely, on segments the model misclassified, this alignment vanished—suggesting those moments truly contain ambiguous or noisy signals.

Taken together, these findings answer our main research question *To what extent can interpretability in AI-based depression diagnosis be improved by identifying salient moments and validating them with humans?.* Overall, our results show that highlighting the exact moments in a patient’s audio-video stream that cause the model’s uncertainty to spike can substantially improve interpretability. When the model correctly flags a segment as “depressed,” the very same facial expressions and vocal cues people observe in those salient moment-long pauses, eye-squints, monotone pitch—are the ones that participants also observe in their own influential moments, and these align neatly with established clinical markers. Moreover, by comparing the radar-chart visualizations, we can directly see which specific features both the model and the participants seem to deem influential in those moments-demonstrating that the cues selected by lay raters (e.g. slowed speech, raised brows, reduced loudness) are not only consistent across viewers, but also potentially truly informative for the model’s decision. In other words, our information-theoretic saliency not only pinpoints **when** the model is being influenced, but also confirms **what** behavioral signals may carry that decision weight in a way that resonates with human intuition and psychological theory. We also observe that the exact timestamps participants select as their ‘tipping point’ often do not coincide with the model’s saliency peaks. Rather than undermining our approach, we interpret this misalignment as a reflection of how difficult it is—even for humans—to pinpoint the precise instant their own judgment changes when viewing very short clips.

References

[1] Akbar, Habibullah et al. (2021). “Exploiting facial action unit in video for recognizing depression using metaheuristic and neural networks”. In: *2021 1st International conference on computer science and artificial intelligence (ICCSAI)*. Vol. 1. IEEE, pp. 438–443.

[2] Al Masud, Gazi Hasan et al. (2025). “Effective depression detection and interpretation: Integrating machine learning, deep learning, language models, and explainable AI”. In: *Array*, p. 100375.

[3] Alghowinem, Sharifa, Roland Goecke, Jeffrey F Cohn, et al. (2015). “Cross-cultural detection of depression from nonverbal behaviour”. In: *2015 11th IEEE International conference and workshops on automatic face and gesture recognition (FG)*. Vol. 1. IEEE, pp. 1–8.

[4] Alghowinem, Sharifa, Roland Goecke, Michael Wagner, et al. (2013). “Eye movement analysis for depression detection”. In: *2013 IEEE International Conference on Image Processing*. IEEE, pp. 4220–4224.

[5] Arik, Sercan O and Tomas Pfister (2021). “Interpretable Deep Learning for Time Series Forecasting”. In: *Google Research*.

[6] Balcombe, Luke and Diego De Leo (2021). “Digital mental health challenges and the horizon ahead for solutions”. In: *JMIR Mental Health* 8.3, e26811.

[7] Baltrušaitis, Tadas, Peter Robinson, and Louis-Philippe Morency (2016). “Openface: an open source facial behavior analysis toolkit”. In: *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, pp. 1–10.

[8] Beck, Aaron T (1996). “Manual for the beck depression inventory-II”. In: *(No Title)*.

[9] Byeon, Haewon (2023). “Advances in machine learning and explainable artificial intelligence for depression prediction”. In: *International Journal of Advanced Computer Science and Applications* 14.6.

[10] Choubey, Hemant and Alpana Pandey (2019). “A new feature extraction and classification mechanisms for EEG signal processing”. In: *Multidimensional Systems and Signal Processing* 30, pp. 1793–1809.

[11] Cummins, Nicholas et al. (2015). “A review of depression and suicide risk assessment using speech analysis”. In: *Speech communication* 71, pp. 10–49.

[12] Ekman, Paul and Wallace V Friesen (1978). “Facial action coding system”. In: *Environmental Psychology & Nonverbal Behavior*.

[13] Ekman, Paul and Erika L Rosenberg (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[14] Fukushima, Kunihiko (1975). “Cognitron: A self-organizing multilayered neural network”. In: *Biological cybernetics* 20.3, pp. 121–136.

[15] Girard, Jeffrey M and Jeffrey F Cohn (2015). “Automated audiovisual depression analysis”. In: *Current opinion in psychology* 4, pp. 75–79.

[16] Girard, Jeffrey M, Jeffrey F Cohn, et al. (2014). “Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses”. In: *Image and vision computing* 32.10, pp. 641–647.

[17] Gomez, Tristan, Thomas Fréour, and Harold Mouchère (2022). “Comparison of attention models and post-hoc explanation methods for embryo stage identification: a case study”. In: *International Conference on Pattern Recognition*. Springer, pp. 216–230.

[18] Gratch, Jonathan et al. (2014). “The distress analysis interview corpus of human and computer interviews.” In: *LREC*. Reykjavik, pp. 3123–3128.

[19] Gu, Wei, Tinghong Ming, and Zhongwen Xie (2023). “Developing a genetic biomarker-based diagnostic model for major depressive disorder using random forests and artificial neural networks”. In: *Combinatorial chemistry & high throughput screening* 26.2, pp. 424–435.

[20] Guidotti, Riccardo et al. (2018). “A survey of methods for explaining black box models”. In: *ACM computing surveys (CSUR)* 51.5, pp. 1–42.

[21] Guntuku, Sharath Chandra et al. (2019). “Understanding and measuring psychological stress using social media”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 13, pp. 214–225.
[22] Gupta, Rahul et al. (2014). “Multimodal prediction of affective dimensions and depression in human-computer interactions”. In: *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pp. 33–40.
[23] He, Lang et al. (2022). “Deep learning for depression recognition with audiovisual cues: A review”. In: *Information Fusion* 80, pp. 56–86.
[24] Hershey, Shawn et al. (2017). “CNN architectures for large-scale audio classification”. In: *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, pp. 131–135.
[25] Hornstein, Silvan et al. (2021). “Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach”. In: *Digital Health* 7, p. 20552076211060659.
[26] Hu, Brian et al. (2023). “Xaitk-saliency: An open source explainable ai toolkit for saliency”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 13, pp. 15760–15766.
[27] Imans, Dillan et al. (2024). “Explainable Multi-Layer Dynamic Ensemble Framework Optimized for Depression Detection and Severity Assessment”. In: *Diagnostics* 14.21, p. 2385.
[28] Jiang, Zifan et al. (2020). “Classifying major depressive disorder and response to deep brain stimulation over time by analyzing facial expressions”. In: *IEEE transactions on biomedical engineering* 68.2, pp. 664–672.
[29] Jin, H et al. (2015). “Predicting depression among patients with diabetes using longitudinal data”. In: *Methods of Information in Medicine* 54.06, pp. 553–559.
[30] Jo, Ashly Ann et al. (2024). “Exploring Explainable AI for Enhanced Depression Prediction in Mental Health”. In: *2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC)*. IEEE, pp. 1–7.
[31] Jordan, Pascal, Meike C Shedden-Mora, and Bernd Löwe (2018). “Predicting suicidal ideation in primary care: An approach to identify easily assessable key variables”. In: *General hospital psychiatry* 51, pp. 106–111.
[32] Joshi, Manju Lata and Nehal Kanoongo (2022). “Depression detection using emotional artificial intelligence and machine learning: A closer review”. In: *Materials Today: Proceedings* 58, pp. 217–226.
[33] Junaid, Sahalu Balarabe et al. (2022). “Recent advancements in emerging technologies for healthcare management systems: a survey”. In: *Healthcare*. Vol. 10. 10. MDPI, p. 1940.
[34] Karimian, Mehran et al. (2025). “A Short Review on Diagnosing and Predicting Mental Disorders with Machine Learning”. In: *International Journal of Applied Data Science in Engineering and Health* 1.1, pp. 20–27.
[35] Karson, CN (1988). “Physiology of normal and abnormal blinking.” In: *Advances in neurology* 49, pp. 25–37.
[36] Kojima, Maki et al. (2002). “Blink rate variability in patients with panic disorder: new trial using audiovisual stimulation”. In: *Psychiatry and clinical neurosciences* 56.5, pp. 545–549.
[37] Kraepelin, Emil (1921). “Manic-Depressive Insanity and Paranoia”. In: *E & S Livingstone*.
[38] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25.
[39] Kroenke, Kurt, Robert L Spitzer, and Janet BW Williams (2001). “The PHQ-9: validity of a brief depression severity measure”. In: *Journal of general internal medicine* 16.9, pp. 606–613.
[40] Kroenke, Kurt, Tara W Strine, et al. (2009). “The PHQ-8 as a measure of current depression in the general population”. In: *Journal of affective disorders* 114.1-3, pp. 163–173.
[41] Kurek, Jarosław, Elżbieta Świderska, and Karol Szymanowski (2024). “Tool Wear Classification in Chipboard Milling Processes Using 1-D CNN and LSTM Based on Sequential Features”. In: *Applied Sciences* 14.11, p. 4730.
[42] LeCun, Yann et al. (1989). “Backpropagation applied to handwritten zip code recognition”. In: *Neural computation* 1.4, pp. 541–551.
[43] Liang, Lijuan et al. (2024). “Enhanced classification and severity prediction of major depressive disorder using acoustic features and machine learning”. In: *Frontiers in Psychiatry* 15, p. 1422020.

[44] Loog, Marco (2011). “Information theoretic preattentive saliency: A closed-form solution”. In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, pp. 1418–1424.
[45] Low, Daniel M, Kate H Bentley, and Satrajit S Ghosh (2020). “Automated assessment of psychiatric disorders using speech: A systematic review”. In: *Laryngoscope investigative otolaryngology* 5.1, pp. 96–116.
[46] Low, Lu-Shih Alex et al. (2010). “Detection of clinical depression in adolescents’ speech during family interactions”. In: *IEEE transactions on biomedical engineering* 58.3, pp. 574–586.
[47] Lucas, Gale M et al. (2015). “Towards an affective interface for assessment of psychological distress”. In: *2015 International Conference on affective Computing and intelligent interaction (ACII)*. IEEE, pp. 539–545.
[48] Lundberg, Scott M and Su-In Lee (2017). “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30.
[49] Mackintosh, JH, R Kumar, and T Kitamura (1983). “Blink rate in psychiatric illness”. In: *The British Journal of Psychiatry* 143.1, pp. 55–57.
[50] Mahayossanunt, Yanisa et al. (2023). “Explainable depression detection based on facial expression using LSTM on attentional intermediate feature fusion with label Smoothing”. In: *Sensors* 23.23, p. 9402.
[51] Mehrabian, Albert (2017). “Communication without words”. In: *Communication theory*. Routledge, pp. 193–200.
[52] Mekonen, Tesfa et al. (2022). “What is the short-term remission rate for people with untreated depression? A systematic review and meta-analysis”. In: *Journal of Affective Disorders* 296, pp. 17–25.
[53] Moreno, Felipe et al. (2023). “Expresso-AI: An Explainable Video-Based Deep Learning Models for Depression Diagnosis”. In: *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp. 1–8.
[54] Mousavian, Marzieh et al. (2021). “Depression detection from sMRI and rs-fMRI images using machine learning”. In: *Journal of Intelligent Information Systems* 57, pp. 395–418.
[55] Mundt, James C, Peter J Snyder, et al. (2007). “Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology”. In: *Journal of neurolinguistics* 20.1, pp. 50–64.
[56] Mundt, James C, Adam P Vogel, et al. (2012). “Vocal acoustic biomarkers of depression severity and treatment response”. In: *Biological psychiatry* 72.7, pp. 580–587.
[57] Näätänen, Risto et al. (2007). “The mismatch negativity (MMN) in basic research of central auditory processing: a review”. In: *Clinical neurophysiology* 118.12, pp. 2544–2590.
[58] Onnela, Jukka-Pekka and Scott L Rauch (2016). “Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health”. In: *Neuropsychopharmacology* 41.7, pp. 1691–1696.
[59] Othmani, Alice, Assaad-Oussama Zeghina, and Muhammad Muzammel (2022). “A model of normality inspired deep learning framework for depression relapse prediction using audiovisual data”. In: *Computer Methods and Programs in Biomedicine* 226, p. 107132.
[60] Parikh, Aditya, Misha Sadeghi, and Bjorn Eskofier (2024). “Exploring Facial Biomarkers for Depression through Temporal Analysis of Action Units”. In: *arXiv preprint arXiv:2407.13753*.
[61] Qi, R et al. (2023). *Explanation strategies for image classification in humans vs. current explainable AI. arXiv*.
[62] Raman, Chirag et al. (2024). “Why Did This Model Forecast This Future? Information-Theoretic Saliency for Counterfactual Explanations of Probabilistic Regression Models”. In: *Advances in Neural Information Processing Systems* 36.
[63] Razykov, Ilya et al. (2012). “The PHQ-9 versus the PHQ-8—is item 9 useful for assessing suicide risk in coronary artery disease patients? Data from the Heart and Soul Study”. In: *Journal of psychosomatic research* 73.3, pp. 163–168.
[64] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). “” Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.
[65] Rother, Anne et al. (2021). “Assessing the difficulty of annotating medical data in crowdworking with help of experiments”. In: *PloS one* 16.7, e0254764.

[66] Scherer, Stefan, Giota Stratou, and Louis-Philippe Morency (2013). “Audiovisual behavior descriptors for depression assessment”. In: *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 135–140.
[67] Schuff, Hendrik et al. (2022). “Human interpretation of saliency-based explanation over text”. In: *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 611–636.
[68] Shah, Varun and Sreedhar Reddy Konda (2021). “Neural Networks and Explainable AI: Bridging the Gap between Models and Interpretability”. In: *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY* 5.2, pp. 163–176.
[69] Sharma, Deepika et al. (2024). “Demystifying Mental Health by Decoding Facial Action Unit Sequences”. In: *Big Data and Cognitive Computing* 8.7, p. 78.
[70] Shatte, Adrian BR, Delyse M Hutchinson, and Samantha J Teague (2019). “Machine learning in mental health: a scoping review of methods and applications”. In: *Psychological medicine* 49.9, pp. 1426–1448.
[71] Sheldon, Kennon M, Mike Corcoran, and Melanie Sheldon (2021). “Duchenne smiles as honest signals of chronic positive mood”. In: *Perspectives on Psychological Science* 16.3, pp. 654–666.
[72] Shin, Daun et al. (2021). “Detection of minor and major depression through voice as a biomarker using machine learning”. In: *Journal of clinical medicine* 10.14, p. 3046.
[73] Siegle, Greg J et al. (2011). “Remission prognosis for cognitive therapy for recurrent depression using the pupil: utility and neural correlates”. In: *Biological psychiatry* 69.8, pp. 726–733.
[74] Slepian, Michael L, Kathleen R Bogart, and Nalini Ambady (2014). “Thin-slice judgments in the clinical context”. In: *Annual review of clinical psychology* 10.1, pp. 131–153.
[75] Song, XinWang et al. (2022). “LSDD-EEGNet: An efficient end-to-end framework for EEG-based depression detection”. In: *Biomedical Signal Processing and Control* 75, p. 103612.
[76] Taguchi, Takaya et al. (2018). “Major depressive disorder discrimination using vocal acoustic features”. In: *Journal of affective disorders* 225, pp. 214–220.
[77] Tang, Yiping et al. (2023). “Video representation learning for temporal action detection using global-local attention”. In: *Pattern Recognition* 134, p. 109135.
[78] Thornicroft, Graham et al. (2017). “Undertreatment of people with major depressive disorder in 21 countries”. In: *The British Journal of Psychiatry* 210.2, pp. 119–124.
[79] Venkataramanan, Kannan and Haresh Rengaraj Rajamohan (2019). “Emotion recognition from speech”. In: *arXiv preprint arXiv:1912.10458*.
[80] Wang, Jingying et al. (2019). “Acoustic differences between healthy and depressed people: a cross-situation study”. In: *BMC psychiatry* 19, pp. 1–12.
[81] Ware, Shweta et al. (2020). “Predicting depressive symptoms using smartphone data”. In: *Smart Health* 15, p. 100093.
[82] Wilson, Sylia and Nathalie M Dumornay (2022). “Rising rates of adolescent depression in the United States: Challenges and opportunities in the 2020s”. In: *Journal of Adolescent Health* 70.3, pp. 354–355.
[83] Woodcock, Claire et al. (2021). “The impact of explanations on layperson trust in artificial intelligence–driven symptom checker apps: experimental study”. In: *Journal of medical Internet research* 23.11, e29386.
[84] World Health Organization (2017). *Depression and other common mental disorders: Global health estimates*. Tech. rep. World Health Organization. URL: https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf.
[85] World Health Organization et al. (2017). “Depression and other common mental disorders: global health estimates”. In.
[86] Yang, Kailai et al. (2023). “Towards interpretable mental health analysis with large language models”. In: *arXiv preprint arXiv:2304.03347*.
[87] Yang, Le et al. (2017). “Hybrid depression classification and estimation from audio video and text information”. In: *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pp. 45–51.
[88] Yang, Ying, Catherine Fairbairn, and Jeffrey F Cohn (2012). “Detecting depression severity from vocal prosody”. In: *IEEE transactions on affective computing* 4.2, pp. 142–150.

[89] Yantis, Steven and John Jonides (1984). “Abrupt visual onsets and selective attention: evidence from visual search.” In: *Journal of Experimental Psychology: Human perception and performance* 10.5, p. 601.
[90] Yasin, Sana et al. (2023). “Machine learning based approaches for clinical and non-clinical depression recognition and depression relapse prediction using audiovisual and EEG modalities: A comprehensive review”. In: *Computers in Biology and Medicine* 159, p. 106741.
[91] Zhang, Linhai et al. (2025). “Explainable Depression Detection in Clinical Interviews with Personalized Retrieval-Augmented Generation”. In: *arXiv preprint arXiv:2503.01315*.
[92] Zhang, Yifei et al. (2023). “XAI benchmark for visual explanation”. In: *arXiv preprint arXiv:2310.08537*.