# Multimodal Emotional Analysis using XAI for Psychotherapy

N. Sripriya[1]; Swetha Subramanian[2]; Sriganesh Jagathisan[3]

[1,2,3]Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, India

Publication Date: 2025/05/23

**Abstract: A person's mental well-being can be perceived by the emotions that they express. What a person feels can be observed by various physical and physiological cues. But people aren't all the same, some are capable of expressing what they truly feel, others might not and there are certain scenarios where the person who is expressing those emotions isn't completely aware of the emotional state they are in. Such scenarios are where even a trained professional isn't always one-hundred percent right. This raises the need for a solution that can observe a person's behavioral traits and guess their emotional state. Currently we have various deep learning approaches that can tackle the problem in hand. One of the widely used approaches is making use of a Unimodal system that predicts a person's emotional state by processing information that is collected in the form of a single modality. But using a single channel to perform such a complex classification task is often inefficient. To make more appropriate classifications, this study proposes a multimodal approach that incorporates eXplainable Artificial Intelligence (XAI) methodologies, and hence improving psychotherapeutic outcomes. The multimodal emotion recognition approach integrates multiple information channels of physical cues of a person, like speech and facial expressions. A more accurate prediction can be arrived at with various complementary channels backing it up. The addition of XAI algorithms make it clearer as to how the model arrived at its conclusion. Overall, this system provides a solution that can be personalized for each client and allows us to have a proper data-driven tool for emotional analysis, which can help the practitioners to design appropriate treatment plans for their client. By adding this state-of-the-art technology as a supplement to conventional psychotherapy techniques, we can yield more successful treatments.**

***Keywords:*** *Multimodal Emotion Recognition; Explainable Artificial Intelligence(XAI); Psychotherapy; Gradcam; LIME; Therapy Results.*

**How to Cite:** N. Sripriya; Swetha Subramanian; Sriganesh Jagathisan. (2025). Multimodal Emotional Analysis using XAI for Psychotherapy. *International Journal of Innovative Science and Research Technology*, 10(5), 1058-1067. https://doi.org/10.38124/ijisrt/25may474.

## I. INTRODUCTION

It is difficult to understand the mental state of human beings by just listening to them talk. First, it is hard for certain people to talk about difficult times in their life and sometimes people can fake their emotions by thinking one thing but saying another. Hence by taking multiple modalities into consideration, in our case, using audio and video, we can predict a person's emotions more efficiently by paying attention to verbal and non-verbal cues. The video input helps us extract facial features and the speech input allows us to perform analysis on various features such as MFCCs, chroma, mel spectrogram, zero crossing rate, RMSE and so on. Using two modalities to predict an emotion of a human is more promising than using a single modality. Moreover, we use XAI – eXplanatory Artificial Intelligence to explain the deep learning model used which seems to be a black-box. XAI provides visual explanations to unravel the black-box. We have used GradCAM for video and LIME for speech to generate heatmaps to explain the features that contribute to the classification process. This is going to be an additional aid to the psychotherapists during their sessions to understand the mental state of their patients in a more effective way. The therapist can be proactive and create better personalized treatment plans.

## II. RELATED WORKS

[1] A study conducted by Khalane et al. in 2023 that focused on context aware multimodal emotion detection using eXplainable Artificial Intelligence (XAI) techniques was proposed. This study has implemented Gradient SHAP to identify key features across audio, video and text modalities in CMU-MOSEI dataset. Their reduced feature models often match or outperform all-feature models and the baseline model GraphMFN. The results proved that selecting these features can reduce noise, bias and improve transparency and trustworthiness of expert systems.

[2] Explainable multimodal learning approaches were studied for engagement analysis of multisensor and multimodal data with continuous performance tests by Rahman et al. (2022). This approach has used Conners' Continuous Performance Test (CPT) to study attention and response inhibition, especially for diagnosing ADHD and neurological disorders. Engagement and disengagements were accurately labelled to monitor cognitive attention. Decision trees and XAI were used for visualization and explanations of the model's prediction.

[3] The research conducted by Guerdan et al. (2021) involves affective explainable AI through facial emotion analysis for understanding human-AI interactions. This approach tries to interpret people's emotional states while interacting with the AI systems by looking at facial action units and arousal levels. It reveals the potential of XAI to interpret the nonverbal cues that are useful to therapists and makes this research relevant to psychotherapy.

[4] Mylona et al. (2022) investigated in a single-session, the alliance breakage and repair processes in psychoanalytic psychotherapy session using multimodal data. Their research highlights how dynamic therapy relationships work and how crucial it is to identity and integrate alliance breaks to achieve good treatment outcomes. This study provides significant details to psychotherapists to improve their treatments by using multimodal analysis with verbal and nonverbal signals of complex interactions.

[5] Terhürne et al. (2022) introduced and validated an automated tool that detects non-verbal emotional expression. Their tool Non-Verbal Behavioral Analyzer (NOVA) seemed to accurately align with conventional practitioner's assessments, as they trained their models over various feature extraction models and found that facial features highly correlated with human-coded arousal ratings. Their study proved that it is possible to make use of automated tools to be used in psychotherapeutic context.

[6] The research conducted by Döllinger et al. (2023) the effect of making trainee psychotherapists undergo standardized computerized emotion recognition accuracy (ERA) training. The trainees were splitted into three groups and were made to undergo different training methods to practice on recognizing emotions. One underwent Multimodal Emotion Recognition Training, where their inputs included audio-only, video-only, and combined audio-video inputs. Then the other group focused mainly on micro expressions to identify concealed emotions. Finally the last cohort was more like an Active control group where they were not given any sort of specialized training. Their study concluded that the trainees that were trained on Multimodal ERA training showed the most promising results, compared to their peers.

[7] Tran et al. (2023) carried out a study which attempts to identify whether the therapeutic outcomes were affected by the empathy that a therapist expressed. By making use of pre-trained models they analyzed the therapist's speech to classify empathy levels. Their model used early and late fusion strategies for their multimodal approach, speech and text inputs. Their observations made it clear that by promoting empathy and rapport building, a therapist can yield better outcomes.

[8] Döllinger et al. (2023) investigated whether the current education curriculum for psychotherapy was improving a trainee's emotion recognition accuracy (ERA). By comparing trainees to a control group of undergraduates, it was observed that initially the trainees had shown promising results compared to the control group. But over the course of 1.5 years there wasn't any improvement from the trainees. This suggested the psychotherapy education wasn't enough to improve ERA, hence emphasizing the need for different trainings and tools to improve ERA.

[9] The MUSE 2022 Multimodal emotion analysis Contest was introduced by Christ et al. (2022) and focused on stress, humor, and emotional reactions. Researchers can benchmark and develop cutting-edge methods for multimodal sentiment analysis on this task. Developments in multimodal methods for sentiment analysis have implications regarding comprehending emotional processes and patterns of communication within therapeutic interactions, even though they are not explicitly related to psychotherapy.

[10] A multimodal sentiment analysis method based on recurrent brain networks and multimodal mechanisms of attention was presented by Cai et al. in 2021. Their research tackles the difficulties involved in interpreting sentiment from a variety of modalities, including spoken word, audio recordings, and visual clues. While their research does not directly address psychotherapy, the creation of strong multimodal sentiment analysis models could improve the understanding and identification of emotions across a range of fields, including psychotherapy.

## III. EXISTING SYSTEM

The current day system used for emotion analysis, be it unimodal or multimodal, aren't usually intuitive in nature. The outcome generated by these models don't speak about what steps made them to draw that outcome as their final labelled outcome, leaving the therapists with only little understanding on why the tool gave them that result. Added to that there are certain situations where the patient may exhibit more than just one emotion at the same time. Instead of identifying such mixed emotions all we get is a singular label stating just one emotion that had the highest probability. Such mixed emotions can be well identified by observing areas where more than one emotion tends to be portrayed. Using XAI we can overcome this difficulty because they can help us visualize the determining factors, for instance using LIME we can determine the parts of a speech input which depicts characteristics similar to a certain type of emotion. So not only do we have a system that's capable of identifying such complex scenarios, but the therapist can also gain insights, identify inaccuracies present in the model and hence tune the model to get the appropriate emotion state that the patient is in. Application of XAI can also help aid to correct, if not overcome, situations where the therapist might have made misinterpretations. Another major concern is the lack of consideration for cultural biases. Using XAI we can overcome this by identifying such cultural cues and weigh them accordingly for the patient in hand. Overall these specified drawbacks in present day techniques for emotion analysis tools can be overcome by the usage of XAI and the multimodal evaluation process can take care of authenticating whether the patient isn't faking his/her emotion, by considering multiple modalities to derive the outcome.

## IV. PROPOSED SYSTEM

The proposed idea aims to come up with an emotional analysis tool by using a hybrid fusion strategy for multiple modalities such as facial and vocal, hence achieving a multimodal system that's capable of predicting the emotional state of a person. The system architecture is detailed in figure.1. With the addition of eXplainbale Artificial Intelligence (XAI) techniques, such as GradCAM and LIME, we improve the model's capability of self-explaining its outcomes. Hence, the therapist will have a means of getting more insight on his patients' emotional well-being and at the same time can give evidence to his patients on why he/she came to that conclusion after their psychological evaluation. Thus, by using the explanations given by the XAI module the therapist can take informed decisions and come up with a tailored treatment plan. Overall, this innovative approach to psychotherapy using a multimodal fusion strategy along with XAI will help us to achieve better personalized medical treatment plans and build trust between the practitioner and the clients.

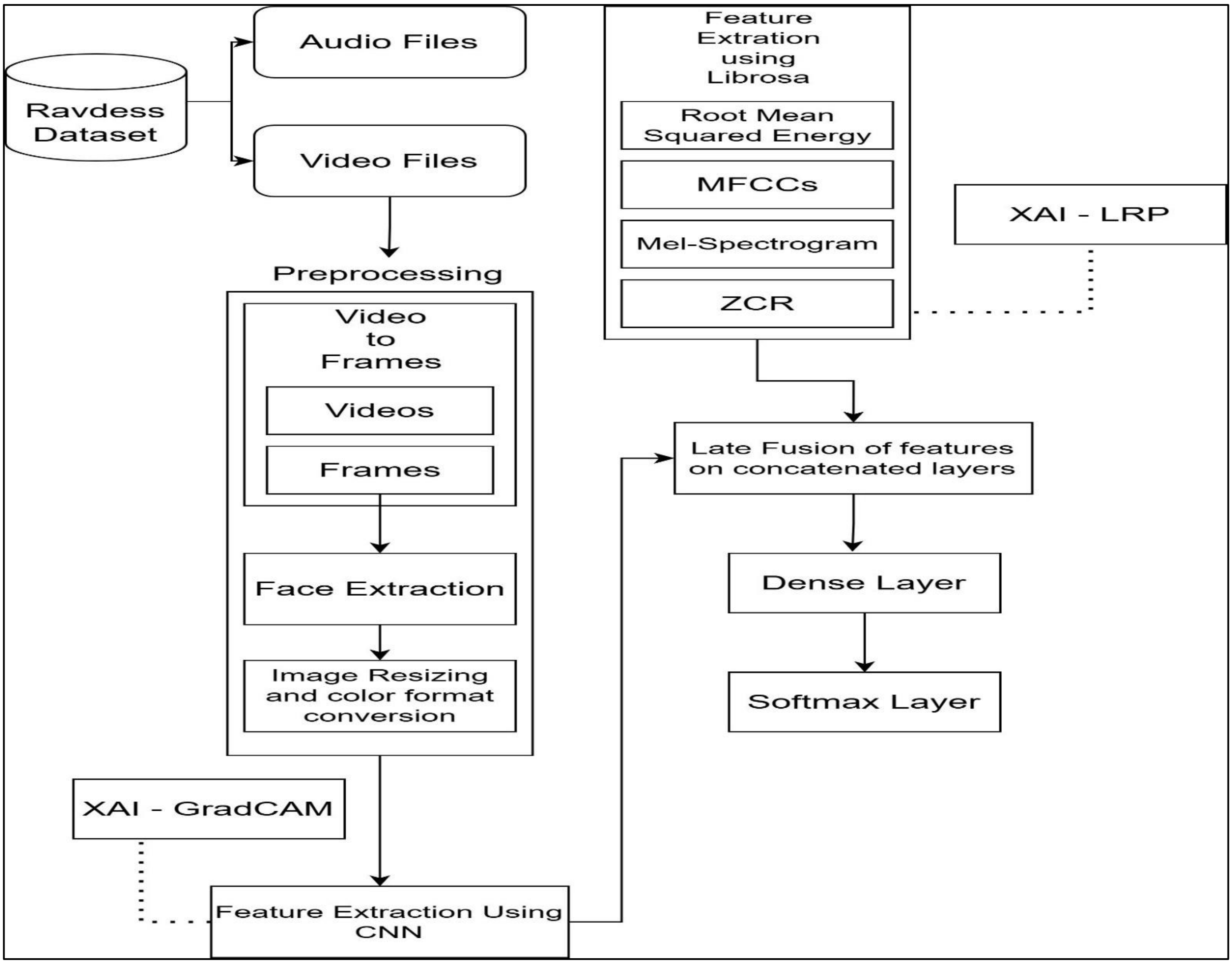

Fig 1: System Architecture

## V. METHODOLOGY

### A. *Data Acquisition Module*

For our task at hand, we chose the Ryerson Audio-Visual Database of Emotions Voice and Song (RAVDESS) dataset. Since it offers a multimodal data with both video and audio recordings of 24 actors portraying a wide range of emotions such as neutral, calm, happy, sad, angry, fearful, disgust, and surprised, we choose this as the most suitable dataset at hand for the use case. Training robust models for deep learning tasks can be done very well using this dataset, due to the availability of a large number of labelled samples per category. Since, diversity of emotions have been depicted appropriately it's widely used and valued for any emotion classification related tasks.

### *B. Data Preprocessing Module*

In this module, we prepared the available dataset inputs in such a way that it makes it easier for us to load the data for the model to ingest and perform its computations as in figure.2.

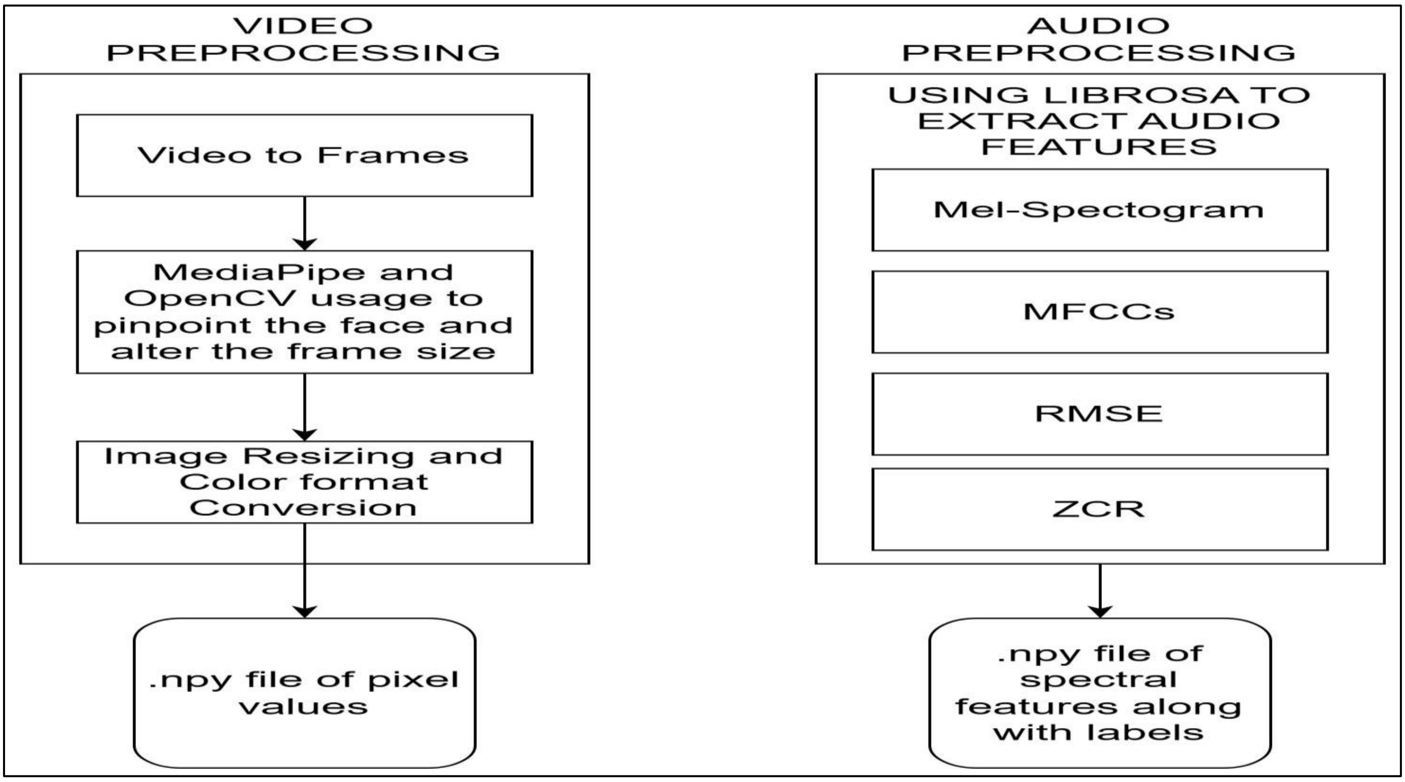
VIDEO PREPROCESSING
Video to Frames
MediaPipe and OpenCV usage to pinpoint the face and alter the frame size
Image Resizing and Color format Conversion
.npy file of pixel values
AUDIO PREPROCESSING
USING LIBROSA TO EXTRACT AUDIO FEATURES
Mel-Spectogram
MFCCs
RMSE
ZCR
.npy file of spectral features along with labels

Fig 2: Input Preprocessing

- For video preprocessing we extracted images frame wise, equally temporally spaced 6 frames, using widely used libraries like MediaPipe and OpenCV. MediaPipe was used to pinpoint the person in the video by fixing points on their facial landmarks.
- For video preprocessing we extracted images frame wise, equally temporally spaced 6 frames, using widely used libraries like MediaPipe and OpenCV. MediaPipe was used to pinpoint the person in the video by fixing points on their facial landmarks.
- Whereas OpenCV was used to resize the frame in such a way that the frame only fitted the person's face, hence removing the unnecessary background whitespace.
- Followed by this a Convolutional Neural Network (CNN) model, trained on image data, was employed to perform the emotion classification on these frames.
- For audio preprocessing, we extracted relevant spectral features like Mel-frequency cepstral coefficients (MFCCs), Mel spectrograms, Zero Crossing Rate (ZCR) and Root Mean Squared Energy (RMSE). These features were identified and fundamental when it comes to making our audio data available in a quantifiable manner.
- All these extracted features were stored in the form of numpy arrays to make data loading easier. This meticulous preprocessing was necessary to transform the available data into a standardized form that can be used for further analysis.

### *C. Hybrid Fusion Module*

Hybrid fusion stands out as an innovative fusion approach for amalgamating insights from both audio and video modalities. By combining information from various stages of each modality, hybrid fusion strategy improves the capability of performing emotion analysis, hence outperforming traditional fusion strategies. This hybrid strategy achieves better accuracy and robustness in terms of psychotherapeutic contexts by capturing complex emotional cues. This is done by a strategic blend of both the audio and visuals features that is extracted at multiple processing stages. When it comes to having a psychotherapy tool that tries to perform emotional analysis via a multimodal approach, hybrid fusion strategy has many advantages to offer compared to the other available fusion strategies. The very first advantage would be making use of information from different modalities, which are complementary, hence allowing the model to make better decisions. This helps the model to identify even the minute changes in emotions, thus improving the model's accuracy. Added to that, this fusion strategy is transparent. Each modalities' contribution to the model's decisions is projected in such a manner that it'll empower the therapists to extract deeper insights about the emotional state of the patient. Thus, the separate audio and visual cues that are available at the therapists' disposal, they can plan on tailoring better interventions. To sum up, the hybrid approach towards multimodality to perform emotion analysis is not only promising to provide more accurate results but also increases the therapeutic outcomes by being both adaptive to each individual's needs.

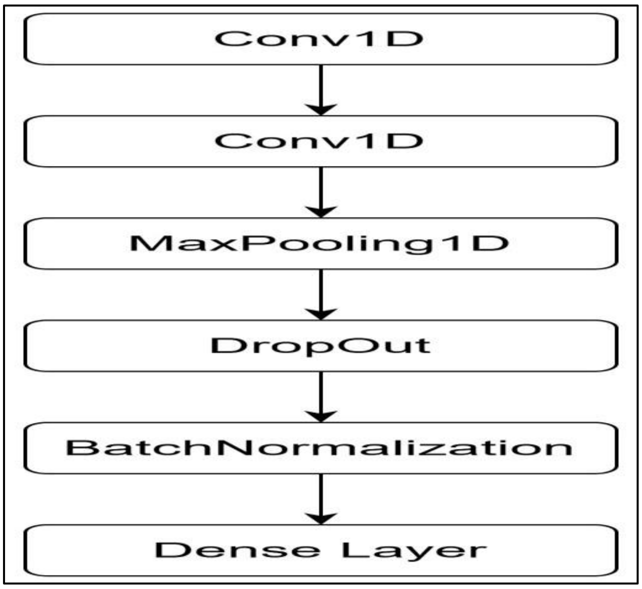

Fig 3: Audio Model Layers

For our implementation we proceeded with fitting a model for each of the modalities, audio and video. For the audio, we used Convolution1D layers, with MaxPooling1D of pool size 8 and a Batch Normalization layer with a 0.4 dropout. The final layer was flattened and given as input to a Dense layer (8) with SoftMax as its activation function, as shown in figure.3. The model was trained and accuracy scores were printed. Similarly, a separate model for video data was created using ConLSTM2D, Convolution2D and MaxPooling2D layers. The pooling layer was flattened and fed to a dense layer (8) with softmax as activation, as shown in figure.4. The model was tested and accuracy scores were printed. For the Fusion model similar structure for the audio and video was created until being passed to their respective dense layers. Then flattened modality data were concatenated and fed to a Dense layer (128) with gelu activation and the final prediction layer was a Dense layer (8) with softmax activation. A model checkpoint was set during training of this fusion model.

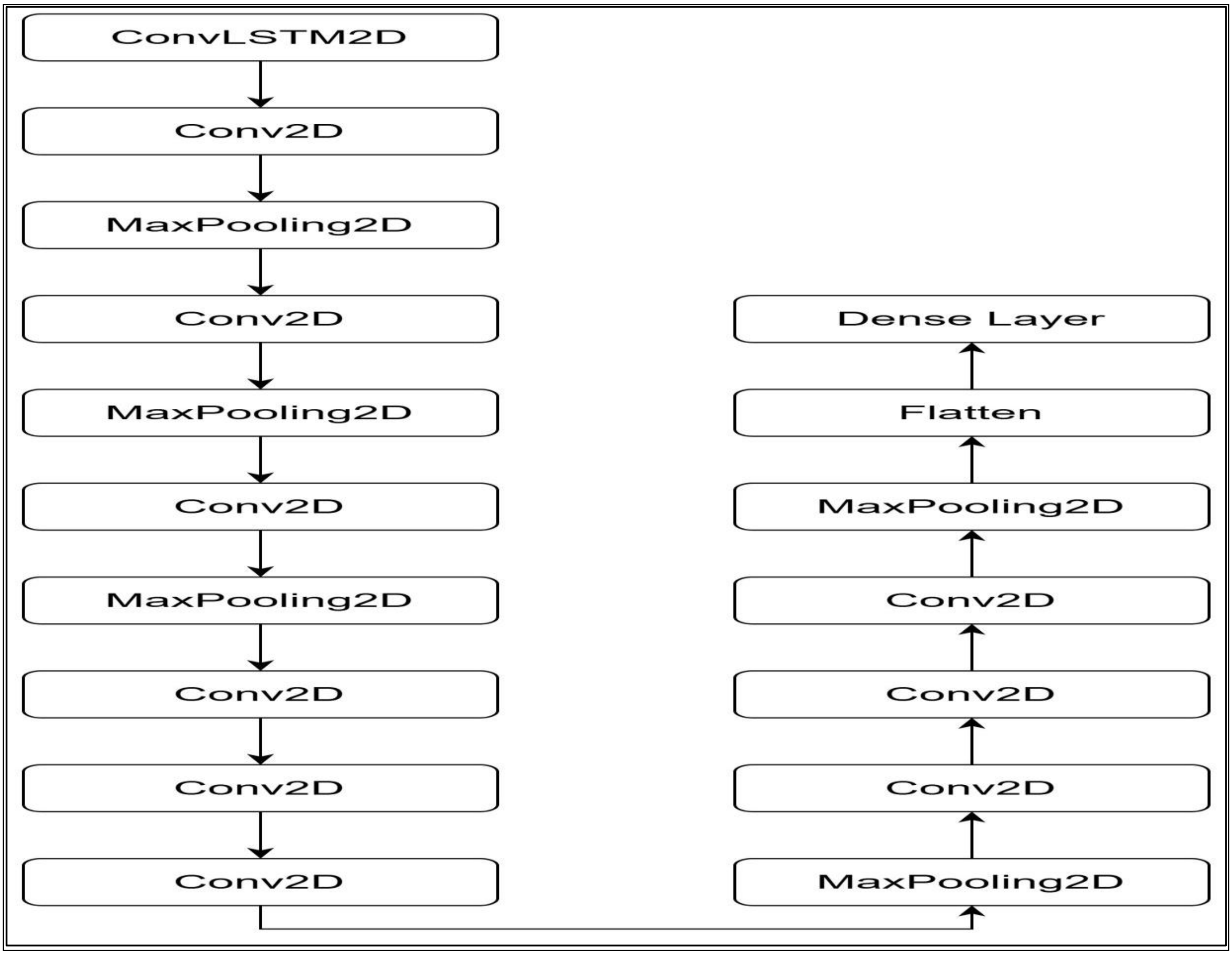

Fig 4: Video Model Layers

### *D. Single Feature Unimodal (Audio) Module*

To compare how the fusion model performs against a unimodal approach we created a separate model for audio inputs which utilized only a single feature, MelSpectrogram. For our matter in hand we used MobileNetV2, figure.5, a pretrained model which is based on CNN architecture. Since it is lightweight and is capable of giving a good amount of performance in terms of computation efficiency, due to depth wise separable convolution, this model was chosen. The generated spectrograms from audio inputs were fed to the pretrained model. Since it's possible to obtain a visual representation of the frequency of audio over time space, mel spectrogram was chosen as the sole feature to be used for this task. The model processes this spectrogram inputs to extract features which are used in out emotion classification task. The Use of this pretrained model also adds to an advantage of obtaining a generalized model for our classification task. This is achieved by the presence of these bottleneck layers in MobileNetv2's architecture. By passing the inputs through a

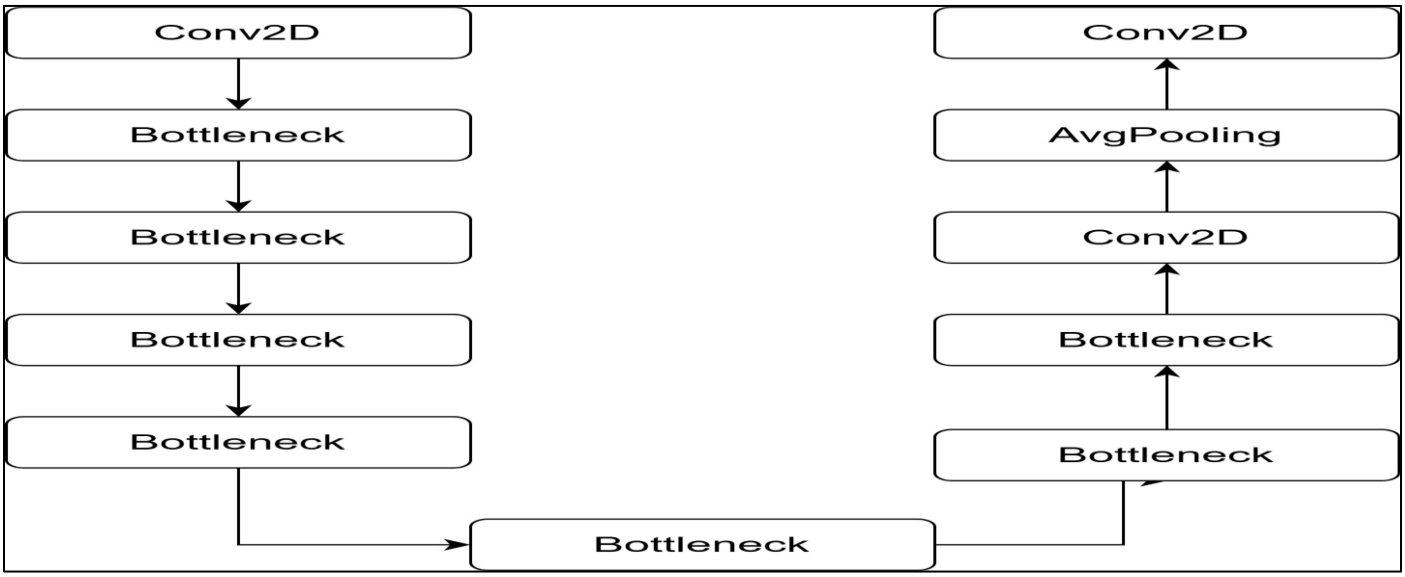

Fig 5: Mobile Net V2 Architecture

Series of lightweight convolution layers, before passing them onto the main convolution layers, the dimensionality of the inputs are reduced. Doing so we can prevent the model from overfitting. Hence, an overall generalized model is obtained. Given that our dataset size is small this comes out to be beneficial. The network is also capable of learning to extract features at multiple scales. Hence, improving the chances of identifying the proper emotions. Considering all these advantages MobileNetV2 has to offer we chose this pretrained model, for our single feature Unimodal model, would be most suited.

### *E. XAI (eXplainable AI) Module*

XAI techniques are used to justify the decision-making process behind predicting the emotion class of each sample. We use the LIME (Local Interpretable Model-agnostic Explanations) to gain insights into the audio features contributing to each predicted emotion class. Additionally, Grad-CAM (Gradient-weighted Class Activation Mapping) and Grad-CAM++ were utilized to generate heat maps highlighting the regions of interest in the input images contributing most to the predicted emotion class.

These XAI techniques enhance interpretability and transparency, facilitating a deeper understanding of the emotion recognition models' decisions. Moreover, they provide valuable insights for psychotherapy, aiding therapists in understanding clients' emotional states and informing more empathetic and effective interventions. The fusion strategy and the XAI techniques used in this study are crucial in psychotherapy for improving the interpretability and robustness of emotion recognition systems.

Table 1 Scores for the Fusion Model

| | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|
| Angry | 0.78 | 0.78 | 0.78 |
| Calm | 0.71 | 0.71 | 0.71 |
| Disgust | 0.79 | 0.50 | 0.61 |
| Fearful | 0.61 | 0.85 | 0.71 |
| Happy | 0.67 | 0.56 | 0.61 |
| Neutral | 0.69 | 0.69 | 0.69 |
| Sad | 0.78 | 0.67 | 0.72 |
| Surprised | 0.61 | 0.89 | 0.72 |
| **Accuracy** | | | **0.69** |

## VI. RESULT AND DISCUSSION

The Multimodal Emotional Evaluation via eXplainable Artificial Intelligence (XAI) tool for psychotherapy uses tone of voice and facial expressions during therapy sessions. The XAI methods used in this tool makes the analysis process transparent and reliable for clients to throw light on how the machine arrived at the results. The application of GradCAM (Gradient-weighted Class Activation Mapping) helped dig deep into the decision-making process of our emotion classification model. By generating the visualizations of salient features within the input images, we identified the significant reasonings behind the model's prediction among different emotion classes. This innovative approach covering different human emotion classes helped to analyze the discriminative facial cues and visual features. The summary of the insights obtained from this approach, as shown in Table.2, reveal the model's decision-making process, giving a deeper understanding of the features and patterns indicative of each human emotion class.

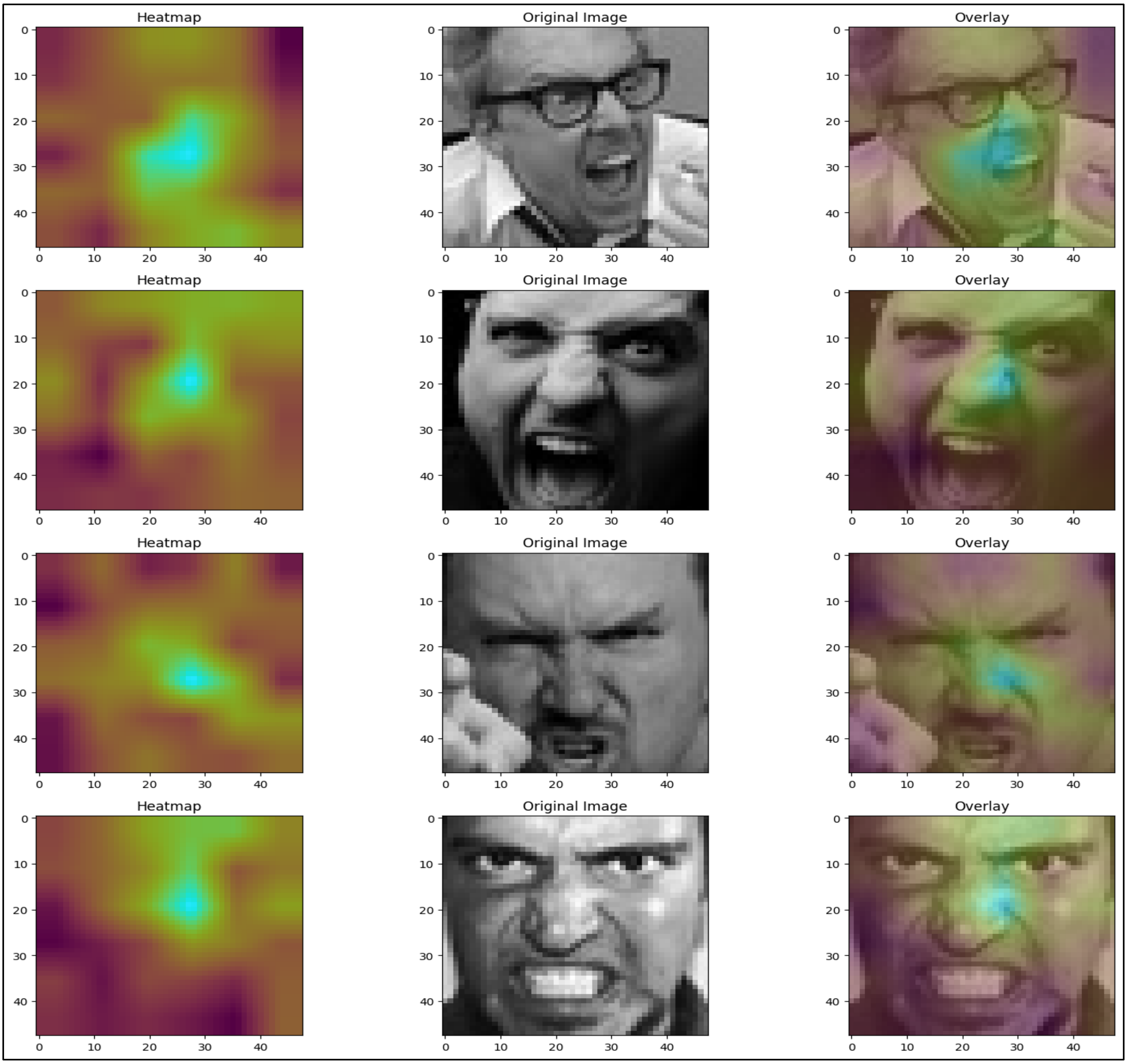

Fig 6: GradCAM - Class Surprised

For the audio inputs, we generated spectrograms and for each spectrogram, we conducted an analysis by iterating through all emotion classes and implementing LIME (Local Interpretable Model-agnostic Explanations) to study the reasoning behind each prediction. The elaborative approach allowed us to gain deeper understanding into the model's decision making process. Visualizing the original spectrogram alongside the LIME heatmaps corresponding to each emotion class, highlighted the specific audio features of the different emotions. In our study, we observed that certain emotions show distinct patterns in their spectrograms, as shown in Table.3. These observations helped us to accurately identify the emotion class. For instance, the neutral emotion seems to display a consistent pattern across various frequency bands and minimal fluctuations in intensity. On the other hand, the calm emotion can be isolated by low-intensity, harmonious patterns indicating a relaxed state.

However, there's potential chances for misclassifications. This type of errors are bound to occur when there are emotions that share similar spectrogram features. A very good example for this scenario would be calm and neutral emotions, both of these emotions exhibit identical characteristics in the bottom green region of the spectrogram. Another possibility for such misclassifications could be due to overlapping spectrogram features. For instance, surprised emotion can be easily mistaken for anger as they share very similar intensity spikes, another example would be sadness and calm emotions, both these have almost same intensity levels across frequency bands. These observations are useful to refine the model, and therefore making it better at classifying the emotion classes.

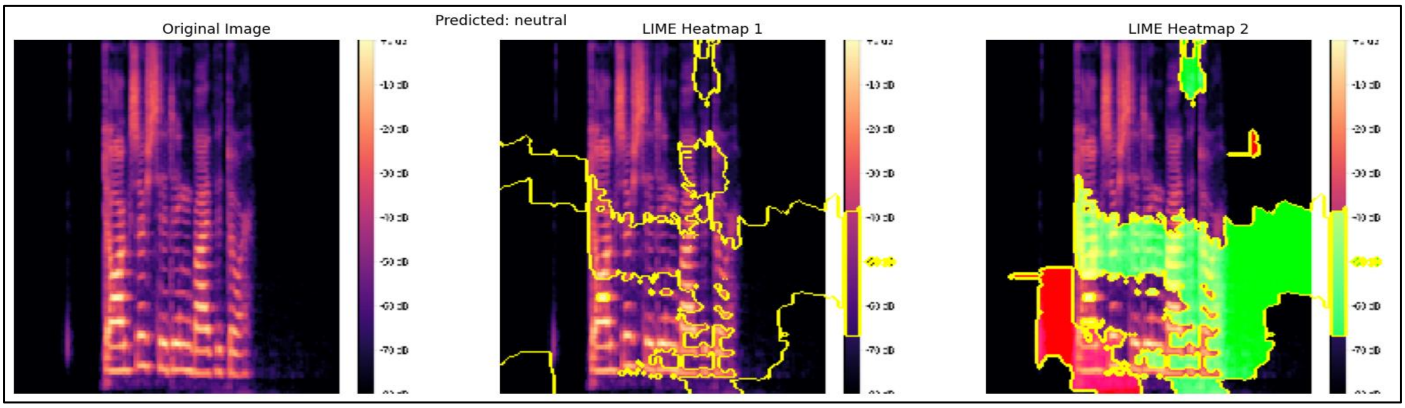

Fig 7: LIME - Class Neutral

By using such tools, psychotherapists can fabricate more personalized and successful treatment plans as they gain significant knowledge about the emotional states of their clients. Moreover, by offering real-time feedback and recommendations in reference to the emotional indicators identified, the tool may make the treatment sessions more fruitful to the clients.

Table 2: Observed Patterns in GradCAM

| Emotions | Table Column Head |
|---|---|
| **Neutral** | Relaxed, even facial expression, Lack of pronounced facial movements or changes |
| | Eyes with a calm, steady gaze, Mouth in a relaxed, closed position |
| | Overall facial features appear balanced and unstressed, Absence of strong emotional cues or expressions |
| **Calm** | Relaxed facial muscles, Soft and gentle facial expressions, Even breathing patters |
| | Even breathing patterns, Mouth in a relaxed, closed position |
| | Neutral or slightly smiling lips, Lack of tension in the face and body |
| **Happy** | Smiling eyes and mouth, Raised cheeks, Bright eyes, Overall expression of joy and contentment |
| **Sadness** | Drooping Eyelids, Downcast Eyes, General expression of sorrow or distress, Lowered Lip comers |
| **Anger** | Lowering of the brow, Narrowing of the lip comers, Eyebrows coming down and together |
| | Raising of the cheekbones, Absence of strong emotional cues or expressions |
| | Facial tension; Clenched Jav, Furrowed brows |
| **Fearful** | Wide eyes, Raised eyebrows, Tense facial muscles |
| | Expression of alarm or distress, Mouth slightly open or gasping |

Considering these results, this approach is a huge step forward for psychotherapy, strengthening emotion analysis boosting treatment results for the patients.

## VII. CONCLUSION

Thus, by combining information from various modalities like audio(speech) and video(facial cues) the

framework shows promise in interpreting emotions accurately as shown in Table.1. With the inclusion of eXplainable artificial intelligence, it becomes clearer on how the final decision was reached, as the model becomes more transparent about its interpretations. This way we can tweak the model's parameters if needed. Therefore, this approach has the potential to provide effective psychotherapy interventions by allowing the therapist to gather meaningful insights from various modal cues, which facilitates in designing the best treatment plans for the client.

Table 3: Observed Patterns in LIME

| Emotion | Inference |
|---|---|
| **Neutral** | Balanced distribution of frequencies with no pronounced peaks or dips |
| **Calm** | Smooth and steady frequency patterns with gentle transitions between different frequency components Might display low-intensity, harmonious patterns indicative of a relaxed state. |
| **Happy** | Could show pronounced peaks in higher frequency ranges. A wide range of frequencies with strong intensity variations, reflecting the positive and enthusiastic nature of happiness. |
| **Sad** | Shows lower overall intensity levels, particularly in the higher frequency bands. Displays a somber and subdued pattern with fewer high-frequency components, reflecting the melancholic tone associated with sadness. |
| **Angry** | Exhibit sharp spikes or peaks in intensity, especially in the mid to high-frequency ranges. Might show abrupt changes in intensity and frequency distribution, reflecting the aggressive and intense nature of anger. |
| **Fearful** | Shows irregular patterns with sudden spikes or drops in intensity across various frequency bands. May show high-frequency components dominating the audio signal, reflecting the tense and anxious state associated with fear. |
| **Disgust** | May display irregular and jagged patterns with fluctuations in intensity across different frequency bins. Might show distorted or dissonant frequency components, reflecting the aversion and repulsion characteristic of disgust. |
| **Surprised** | Reflected by sudden and intense peaks in intensity across a wide range of frequencies. Might exhibit sharp and unexpected changes in frequency distribution, reflecting the abrupt and startled nature of surprise. |

## FUTURE WORK

For future enhancements of the proposed system, there is huge scope when it comes to using Explainable Artificial Intelligence (XAI), by exploring various options to enhance the capability of the model to explain its decisions. Making the explanatory forms more easily interpretable could also be an area of research. By providing a visual means that's simpler and yet properly covering the model's interpretations would not only improve the confidence of the therapists over the model but would also make it convenient to work with. Another area for future scope would be to make the system collaborate well with its user, that is, by giving more flexibility to align the predictions of the model with that of the therapist's expertise. Additionally, there's scope for emotion validation which can be achieved by incorporating more modalities especially physiological ones. With addition of physiological modalities like ECG we can confirm if the emotion depicted is authentic and not fake. A good example would be checking if a patient is purposefully depicted as if he is calm in an anger management session. Since it is quite possible to fake most of the physical modalities, usage of physiological features, such as EEG or ECG, which are generally used in most therapy sessions to monitor whether a person is being truthful during that session. Such a pattern analysis would give more confidence over the tool for both the client and the practitioner. One more place for scope would be to try different fusion strategies. Using early, late or intermediate fusion strategies on the same set of modalities can be used to identify which among them is a better fusion strategy. Mix and match of multiple modalities with intermediate fusion might bring out better model performance in terms of computation costs. Finally, evaluating the long-term efficacy and benefits of utilizing such a system in real-world psychotherapy settings could be a valuable direction for future studies.

## REFERENCES

[1]. Khalane, A., Makwana, R., Shaikh, T., & Ullah, A. (2023). Evaluating significant features in context-aware multimodal emotion recognition with XAI methods. Expert Systems, e13403.

[2]. Rahman, M. A., Brown, D. J., Shopland, N., Burton, A., & Mahmud, M. (2022, June). Explainable multimodal machine learning for engagement analysis by continuous performance test. In International Conference on Human Computer Interaction (pp. 386-399). Cham: Springer International Publishing.

[3]. Guerdan, L., Raymond, A., & Gunes, H. (2021). Toward affective XAI: facial affect analysis for understanding explainable human-ai interactions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3796-3805).

[4]. Mylona, A., Avdi, E., & Paraskevopoulos, E. (2022). Alliance rupture and repair processes in psychoanalytic psychotherapy: multimodal in-session shifts from momentary failure to repair. Counselling Psychology Quarterly, 35(4), 814-841.

[5]. Terhürne, P., Schwartz, B., Baur, T., Schiller, D., & André, E. (2022). Validation and application of the Non Verbal Behavior Analyzer: An automated tool to assess non verbal emotional expressions in psychotherapy. Frontiers in Psychiatry, 13, 1026015.

[6]. Döllinger, L., Högman, L. B., Laukka, P., Bänziger, T., Makower, I., Fischer, H., & Hau, S. (2023). Trainee psychotherapists' emotion recognition accuracy improves after training: emotion recognition training as a tool for psychotherapy education. Frontiers in Psychology, 14, 1188634.

[7]. Tran, T., Yin, Y., Tavabi, L., Delacruz, J., Borsari, B., Woolley, J. D., ... & Soleymani, M. (2023, October). Multimodal Analysis and Assessment of Therapist Empathy in Motivational Interviews. In Proceedings of the 25th International Conference on Multimodal Interaction (pp. 406-415).

[8]. Döllinger, L., Letellier, I., Högman, L., Laukka, P., Fischer, H., & Hau, S. (2023). Trainee psychotherapists' emotion recognition accuracy during 1.5 years of psychotherapy education compared to a control group: no improvement after psychotherapy training. PeerJ, 11, e16235.

[9]. Christ, L., Amiriparian, S., Baird, A., Tzirakis, P., Kathan, A., Müller, N., ... & Schuller, B. W. (2022, October). The muse 2022 multimodal sentiment analysis challenge: humor, emotional reactions, and stress. In Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge (pp. 5-14).Terhürne, P., Schwartz, B., Baur, T., Schiller, D., & André, E. (2022). Validation and application of the Non Verbal Behavior Analyzer: An automated tool to assess non verbal emotional expressions in psychotherapy. Frontiers in Psychiatry, 13, 1026015.

[10]. Cai, C., He, Y., Sun, L., Lian, Z., Liu, B., Tao, J., ... & Wang, K. (2021). Multimodal sentiment analysis based on recurrent neural network and multimodal attention. In Proceedings of the 2nd on multimodal sentiment analysis challenge (pp. 61-67).