Attention-Optimized Fusion of Multiple Data Modalities for Psychological Disorder Assessment

Slah Rabaoui[2], Samar Bouazizi[1] [0000-0002-8793-1128] and Hela Ltifi[2] [0000-0003-3953-1135]

[1]Research Groups in Intelligent Machines Lab, BP 3038, Sfax, Tunisia
[2]Faculty of Sciences and Technology of Sidi Bouzid, University of Kairouan, Tunisia
rb.slah1@gmail.com,samar.bouazizi@enis.tn, hela.ltifi@ieee.org

Abstract. **Detection of mental health conditions like anxiety, depression, and post-traumatic stress disorder (PTSD) at their early stages is crucial for successful treatment interventions. This paper introduces an innovative framework that combines multiple data modalities for identifying psychological disorders. Our approach synthesizes three distinct data sources: audio recordings of speech, text transcriptions of conversations, and clinical measurements including PHQ-8 questionnaire results and patient demographics. At the core of our methodology is a sophisticated attention-driven fusion system that intelligently calibrates the influence of each data stream according to individual patient contexts, generating a comprehensive representation of their mental state. To support clinical understanding, we implement explainable artificial intelligence methodologies (SHAP) that highlight key contributing factors in the classification process and offer healthcare providers meaningful insights into the model's decision-making. The system demonstrated exceptional performance with 95.83% accuracy in differentiating between anxiety, depression, and PTSD cases, surpassing previous approaches that relied on single data types.**

Keywords: **Multimodal approach, attention-based fusion, mental health assessment, anxiety, depression, PTSD, clinical decision support,**

I. INTRODUCTION

Mental health disorders represent a major global health challenge, affecting nearly one billion people worldwide [1], with anxiety, depression, and post-traumatic stress disorder (PTSD) being particularly widespread and debilitating. Several research has confirmed the validity of vocal biomarkers for psychiatric diagnosis, with studies suggesting that speech characteristics like monotony, reduced prosodic variation, and disrupted speech rates can discriminate depression with more than 80% accuracy [2, 9, 10]. Similarly, some voice patterns have been associated with PTSD, including disrupted pitch variability and elevated speech disfluencies [3]. These studies have noted increased use of first-person pronouns and affective negative words in depression [4] and, on the other hand, unnecessary hedges and apprehension regarding the future in anxiety disorders [5]. Traditional clinical tools like PHQ-8 provide standardized measurement with established validity [6], yet the heterogeneity of psychological disorders means that there is no single data source capable of capturing their complexity. Our research introduces a novel attention-based fusion mechanism that learns to balance each modality's contribution dynamically based on patient-specific contexts. By incorporating transformer architectures, which excel at processing sequential data with complex temporal dynamics, our approach captures the prosodic patterns and semantic structures that signal psychological distress. In the following sections, we present a theoretical background, next, we detail our methodology, including data preprocessing, feature extraction, model architecture, and evaluation approach. We then present our results, highlighting the performance advantages of our multimodal approach compared to unimodal baselines. Finally, we conclude our findings and outline directions for future research in this rapidly evolving field.

II. THEORETICAL BACKGROUND

To overcome the early detection of depression and anxiety, [7] created a model integrating machine learning techniques with psychological questionnaires. Two datasets have been used in their method to classify symptom severity using CNN, SVM, LDA, and K-NN. Among these, CNN was best with 96% accuracy for anxiety and 96.8% for depression. SVM also performed equally well (95% and 95.8% in anxiety and depression, respectively). LDA did the next better (93% and 87.9%, respectively). The performance of K-NN was poorer with an accuracy of 70.96% (anxiety) and 81.82% (depression) respectively, reflecting the applicability of CNN as a clinical tool for treatment individualization.[8] applied binary SVM to separate generalized anxiety disorder (GAD) and major depressive disorder from healthy controls using multimodal data, e.g., cortisol levels and gray matter volume. While clinical questionnaires by themselves had limitations in GAD prediction, the integration of biological markers raised the level of accuracy to 90.10% for case identification and 67.46% for differentiation of the disorder, thereby highlighting the necessity for integrative data approaches. Supporting these results, [9] contrasted text and audio data for depression identification with emphasis on feature enhancement and F1-score optimization. Then, [11] proposed a random forest classifier with enhanced band-pass filtering for stress classification from ECG signals. They trained on the MIT-BIH database (47 patients) and tested on real-time data (11 patients) and achieved 96.73% accuracy in stress level classification with consistent real-time performance.

III. METHODOLOGY

Our research adopts a comprehensive approach to psychological disorder detection by analyzing multiple types of patient data simultaneously (cf. Fig.1). We first pre-process every type of data individually. In the case of audio recordings, we clean the sound files and extract features that might indicate psychological distress, such as changes in voice tone, rhythm, and quality. It has been established in several studies that depression and anxiety often manifest in speaking patterns, with depressed individuals typically speaking more slowly and monotonously. For text data, we analyze the recorded speech, such as word choice, affective content, frequency of personal pronouns, and general use of language. Clinical information including questionnaire responses and demographic details provides additional context. In the second step, we integrate these various data sources. This requires careful alignment to ensure that information from the same patient is properly matched across all sources. We also handle missing data and standardize all of the variables in order to be able to compare them. By bringing together audio, text, and clinical data, we create a more complete picture of each patient than any single data type could provide. The third step involves selecting the most relevant features from our combined dataset. Not all measurable aspects contribute equally to psychological assessment, so we identify the most informative indicators from each data type. This focused approach improves the efficiency and accuracy of our model by reducing noise and redundancy. The last step implements our innovative multimodal model.

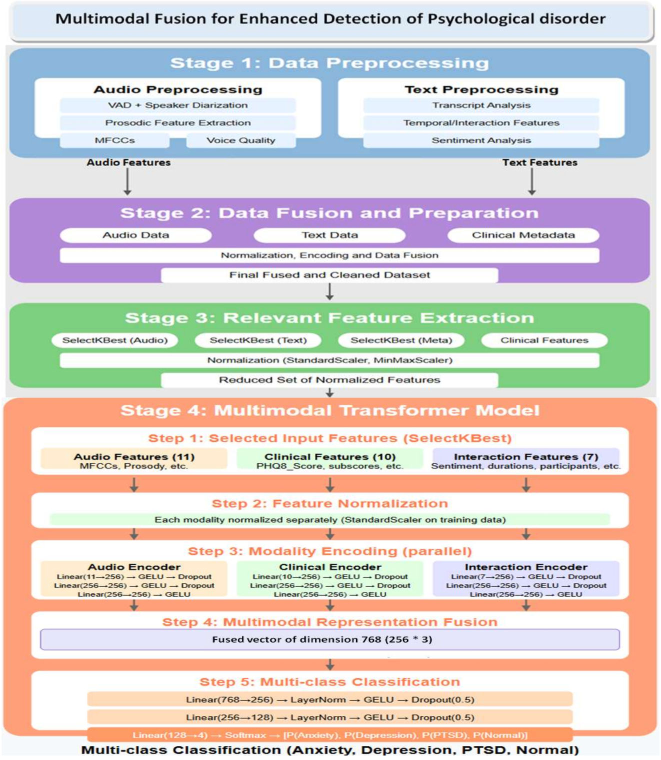

Multimodal Fusion for Enhanced Detection of Psychological disorder
Stage 1: Data Preprocessing
Audio Preprocessing
VAD + Speaker Diarization
Prosodic Feature Extraction
MFCCs
Voice Quality
Text Preprocessing
Transcript Analysis
Temporal/Interaction Features
Sentiment Analysis
Audio Features
Text Features
Stage 2: Data Fusion and Preparation
Audio Data
Text Data
Clinical Metadata
Normalization, Encoding and Data Fusion
Final Fused and Cleaned Dataset
Stage 3: Relevant Feature Extraction
SelectKBest (Audio)
SelectKBest (Text)
SelectKBest (Meta)
Clinical Features
Normalization (StandardScaler, MinMaxScaler)
Reduced Set of Normalized Features
Stage 4: Multimodal Transformer Model
Step 1: Selected Input Features (SelectKBest)
Audio Features (11)
MFCCs, Prosody, etc.
Clinical Features (10)
PHQ8_Score, subscores, etc.
Interaction Features (7)
Sentiment, durations, participants, etc.
Step 2: Feature Normalization
Each modality normalized separately (StandardScaler on training data)
Step 3: Modality Encoding (parallel)
Audio Encoder
Linear(11→256) → GELU → Dropout
Linear(256→256) → GELU → Dropout
Linear(256→256) → GELU
Clinical Encoder
Linear(10→256) → GELU → Dropout
Linear(256→256) → GELU → Dropout
Linear(256→256) → GELU
Interaction Encoder
Linear(7→256) → GELU → Dropout
Linear(256→256) → GELU → Dropout
Linear(256→256) → GELU
Step 4: Multimodal Representation Fusion
Fused vector of dimension 768 (256 * 3)
Step 5: Multi-class Classification
Linear(768→256) → LayerNorm → GELU → Dropout(0.5)
Linear(256→128) → LayerNorm → GELU → Dropout(0.5)
Linear(128→4) → Softmax → [P(Anxiety), P(Depression), P(PTSD), P(Normal)]
Multi-class Classification (Anxiety, Depression, PTSD, Normal)

Fig.1. Multimodal Fusion for Enhanced Detection of Psychological Disorders

In contrast to previous methods that might possibly process voice or text separately, all the data categories are handled by our model simultaneously. Our model identifies how various features present together and what dynamic weights of importance each has, taking into account that certain signs are appropriate for specific patients but not others. This replicates how human doctors compound more than one observation to form an opinion. Our inspiration for developing this multimodal approach is the high internationality of psychiatric disorder burden and the long-standing difficulty with their early and accurate detection. A critical component of our methodology is the incorporation of Explainable AI (XAI) techniques to enhance the interpretability and transparency of our deep learning model.

1. Used dataset

The dataset used in our study was the Distress Analysis Interview Corpus with Wizard-of-Oz (DAIC-WOZ), developed by the University of Southern California (USC) Institute for Creative Technologies. The DAIC-WOZ corpus consists of 189 clinical interview recordings and transcripts of participants' interactions with a virtual agent named "Ellie" operated by human controllers. Each session has audio recordings, text transcripts, and meta clinical data like PHQ-8 depression screening scores (0-24, higher than 10 indicating depression) and demographic information like gender. The data are divided into 107 for training, 35 for validation, and 47 for testing (with 33 non-depressed and 14 depressed subjects in the test set). The interviews were constructed to release indicators of psychological distress and were analyzed by modality from the patients' answers to determine markers of depression, post-traumatic stress disorder, and anxiety. They are stored in individual folders labeled 300_P to 492_P, but some sessions (342, 394, 398, 460) were not used because they were not recorded appropriately, and some have breaks in between. This large multimodal corpus provides essential data for the training and testing of AI models for psychological condition detection through verbal and non-verbal cues.

2. Data preprocessing

The preprocessing pipeline consistently transforms raw multimodal inputs: into harmonized, structured forms ensuring robust input to the multimodal deep learning model.

➢ ***Audio Data Preprocessing:*** The audio pre-processing method follows a comprehensive multi-step procedure dedicated to the extraction of informative acoustic features from patient voice recordings. To begin, each audio file is normalized so that amplitude levels are equalized, followed by noise reduction that eliminates background noise potentially contaminating the signal. The system subsequently uses automated speaker diarization and voice activity detection (VAD) in order to split speech regions from silence periods and detect various speakers. This segmentation, performed through energy-based signal analysis, enables the separation of critical temporal measurements like speech duration, pause duration, speech rate, and articulation rate—all valuable measures of speech fluency and rhythm. The acoustic feature extraction pipeline subsequently transforms raw vocal signals into a structured set of 47 measurable parameters. Among them are Mel-

Frequency Cepstral Coefficients (MFCCs), which represent the envelope-like patterns in the spectra on the mel scale to more accurately reflect human hearing. Voice quality measurements such as jitter (frequency variation), shimmer (amplitude variation), and Harmonics-to-Noise Ratio (HNR) are computed to estimate vocal stability and clarity—apt here since unstable or monotone speech can express psychological distress. Spectral features like spectral energy, bandwidth, and centroid are calculated using Fast Fourier Transform (FFT) to analyze frequency distribution patterns in speech. Each patient is given a unique identifier (Patient_id) for tracing the data through the analysis to allow correspondence between extracted features from the voice and clinical diagnoses. This systematic process of vocal data conversion results in a dense vectorial form that enables deep learning algorithms to recognize nuanced acoustic differences that might be characteristic of different psychological states.

➢ ***Text Data Preprocessing:*** The textual preprocessing pipeline transforms raw interview transcriptions into structured, machine-learning-ready representations through several sophisticated stages. The process begins with loading the text files, followed by character normalization, including lowercase conversion for normalizing the data and suppressing unnecessary differences between caseless and cased letters. A clean step removes extraneous components like punctuation, special characters, and numbers that contribute minimal semantic meaning to the analysis. Tokenization—the very crucial preprocessing operation—splits text into elementary units known as tokens (words or subwords), allowing intricate analysis of language patterns. Stop word filtering eliminates common words with minimal or zero semantic value (e.g., "the," "and," "is"), thereby reducing noise and dimensionality from the data. Lemmatization or stemming techniques reduce words to their canonical or root form, consolidating different morphological forms of the same word into a common representation. These are followed by the system extracting higher-level linguistic features to assess the psychological state of the patient. Lexical richness analysis identifies potential vocabulary poverty, which is one potential indicator for depressive or cognitive disorders. Syntactic structure analysis examines sentence complexity, insofar that short, fragmented sentences may be signs of reasoning disorders or thought disorder. The system monitors repetition and frequency of negatively charged words (e.g., "sad," "tired," "alone") and first-person pronoun use (e.g., "I," "me"), potential markers for psychological distress or isolation. Sentiment analysis takes the proportion of positive to negative words in patient language, which helps detect depressive or anxious propensity. High-end models like word embeddings (Word2Vec, GloVe, BERT) map semantic word relationships in a way that cognitive thought patterns characteristic of specific psychological conditions become recognizable. Vectorized word representations are input into a Transformer-based deep learning architecture, and it becomes feasible to accurately classify the psychological state of the patient through analysis of multiple dimensions of language.

➢ ***Meta clinical Data Preprocessing:*** The meta clinical data preprocessing method is directed towards contextualizing audio and text analyses through comprehensive combination of survey responses and demographic information. The process begins by standardizing PHQ-8 (Patient Health Questionnaire) scores, which is a depression screen validated with scores above 10 being indicative of likely depression. Each of the eight individual question responses is normalized and inspected independently, providing granular data concerning discrete symptom domains such as sleep disturbance, energy, concentration difficulties, and anhedonia (loss of interest or pleasure). The individual question analyses provide more pattern detection detail than the total score. Demographic variables are encoded with care—age values are normalized to prevent scale-based model bias, and gender is converted to numeric using one-hot encoding to maintain categorical integrity. When available, medical history data like previous depressive episodes and medication regimens are standardized and encoded to provide additional clinical context. Median imputation is employed for missing value treatment of the meta clinical data in a manner that retains distributional characteristics of the data and does not reduce sample size. Correlation analysis is employed to detect and treat multicollinearity in clinical variables for model stability improvement. The meta clinical features are preprocessed and aligned with the corresponding audio and text data based on patient identifiers, ensuring accurate temporal and contextual alignment across modalities. This comprehensive meta clinical data preprocessing approach enables the model to incorporate useful clinical and demographic contexts in interpreting speech patterns and linguistic features, significantly enhancing the accuracy and interpretability of psychological state classification.

3. Data fusion

The multimodal data fusion approach for psychological disorder detection involves a sophisticated, multi-stage process that harmonizes diverse data types into a cohesive representation suitable for transformer-based classification. First, three distinct streams of data: audio recordings, text transcriptions, and clinical metadata, are preprocessed separately with domain-specific techniques: audio data are preprocessed using Voice Activity Detection (VAD), speaker diarization, and acoustic feature extraction like MFCCs, prosodic features, and voice quality measures using Librosa and COVAREP; textual data are processed using transcript analysis, temporal/interaction feature extraction, and sentiment analysis with BERT vectorization; and clinical metadata with PHQ-8 scores and demographics are normalized. The merging process begins by synchronizing these heterogeneous sources through the Patient_id key so that only complete patient profiles with information regarding all modalities are retained, thus eliminating training bias as a result of incomplete information. In the second stage of data preparation, the fusion process addresses several important challenges: text interaction counts in the form of dictionaries are transformed into separate numerical columns (retaining only the discriminative ellie_count variable and discarding the non-discriminative participant_count); missing values in acoustic and textual features are imputed with median-based methods to preserve data structure while minimizing the effect of outliers; categorical variables like Gender are encoded numerically (Male/Female to 0/1); and all numeric features are normalized to zero-mean, unit-variance distributions to prevent disproportionate influence from variables with larger ranges. The third step utilizes feature optimization by variance analysis to remove low-variation and non-discriminative features and correlation analysis to remove redundant features (features with correlations >0.95), thereby reducing dimensionality without compromising information integrity. Finally, the combined dataset is reordered with Patient_id as the first column and sorted in ascending order to ensure traceability before being fed into the multimodal transformer model, which makes use of an intricate attention mechanism that dynamically dictates the weight given to the contribution of each modality depending on its applicability for each patient's psychological state classification.

4. Feature extraction

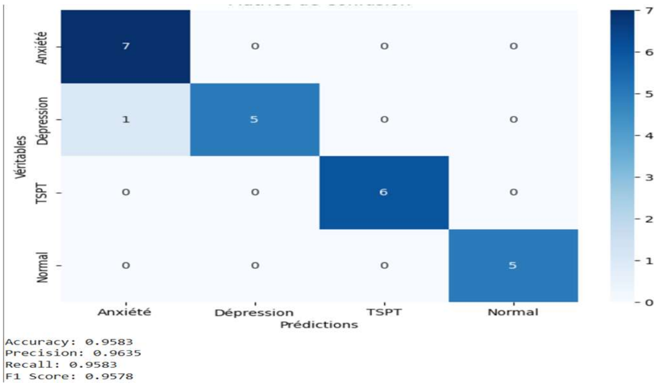
In the feature extraction step we used Fast Fourier Transform (FFT) as the backbone in obtaining spectral features from DAIC-WOZ audio signals. In the initial step of audio preprocessing, raw waveforms are windowed using Hamming windows of width 25ms and frame shift 10ms to minimize spectral leakage. The FFT algorithm, with a default frame size of 512 points, subsequently transforms these time-domain segments into frequency-domain representations, which reflect the distribution of energy over different frequency bands. This spectral decomposition enables the derivation of important acoustic features like Mel-Frequency Cepstral Coefficients (MFCCs), which are computed by passing the power spectrum through a Mel filterbank, followed by logarithmic compression and Discrete Cosine Transform (DCT). The system uses 13 principal MFCCs and their first and second derivatives (delta and delta-delta coefficients) to represent vocal tract dynamics. Furthermore, FFT facilitates the measurement of prosodic parameters like the fundamental frequency (F0) via autocorrelation in the frequency domain, spectral centroids for the "center of mass" of the spectrum, spectral flux for frame-to-frame spectral rate of change calculations, and spectral flatness to quantify the quality of tones compared to noise-like character of the speech signal. Voice quality measures such as jitter (pitch period perturbations) and shimmer (amplitude perturbations) are computed from spectral peaks measured via FFT, and Harmonic-to-Noise Ratio (HNR) is obtained by ratioing the energy of the harmonic component to that of the noise floor in the spectrum. These FFT-extracted acoustic features are normalized and standardized before the SelectKBest method identifies the most discriminative acoustic parameters in creating a rich representation of voice features which may be the indicative features for quantifying psychological distress, particularly the differentiation of anxiety, depression, and post-traumatic stress disorder under the multimodal classification scenario.

5. Multimodal transformer

The multimodal transformer model architecture is a sophisticated solution for classification of psychological disorders that is organized into five steps of processing. The model initiates with feature selection (Step 1), whereby it strategically processes 11 audio features (MFCCs and prosody metrics), 10 clinical features (PHQ8 scores and subscores), and 7 interaction features (sentiment analysis, speech durations, and participant statistics). In Step 2, each modality is individually normalized by StandardScaler trained on the training data to equalize the contribution of features regardless of their original scales. The core of the architecture is in Step 3, wherein parallel modality encoding is realized through committed neural pathways. The Audio Encoder transforms the 11 input features through a sequence of linear layers (11→256→256→256) with GELU activations and strategic dropout between layers for regularization. The Clinical Encoder acts on 10 features and the Interaction Encoder on 7 features, both with the same structures that terminate at 256-dimensional representations. This parallel processing allows each modality to create expert representations before fusion. Step 4 employs multimodal representation fusion, which combines the three 256-dimensional vectors into a single 768-dimensional fused representation (256 × 3). This critical fusion step allows the model to capture cross-modal interactions and dependencies between audio cues, clinical markers, and interaction patterns. The final classification step (Step 5) employs a sequence of fully connected layers that progressively reduce dimensionality (768→256→128→4) with layer normalization, GELU activations, and 0.5 dropout probabilities to prevent overfitting. The output layer uses softmax activation to produce probability distributions across four classes: Anxiety, Depression, PTSD, and Normal. The model is trained using the Adam optimizer and dynamic learning rate scheduling, batch size 32, and a maximum of 50 epochs with early stopping to prevent overfitting. Weighted cross-entropy loss is employed to combat class imbalance, while data augmentation techniques augment the audio samples. The model's strength lies primarily in its multimodal approach, which acknowledges the multifaceted nature of psychological disorders. By integrating audio features (capturing vocal biomarkers like prosody and MFCCs), clinical data (PHQ-8 scores), and interaction patterns (sentiment and conversational dynamics), the architecture addresses the limitations of unimodal approaches that fail to capture the complex manifestations of mental health conditions. This integration reflects clinical reality, where practitioners rely on multiple indicators across different modalities to make accurate diagnoses. The consistent use of 256-dimensional embeddings across all three encoders creates balanced representations that contribute equally to the final fusion, avoiding modality bias.

IV. Experimental results

The confusion matrix provides a full visualization of the multimodal transformer model's discriminative power in classifying among the four states of mind. With an overall accuracy recorded at 95.83%, the model demonstrates excellent discriminative ability. The model displayed excellent precision (96.35%) and recall (95.83%), resulting in an F1 score of 95.78% that confirms the overall balanced effectiveness for all the classes. Our multimodal transformer model worked excellently in identifying psychological disorders through extensive 5-fold cross-validation. The model achieved 95.83% accuracy in distinguishing among anxiety, depression, PTSD, and normal states, which was highly better than single-modality approaches (audio-only: 83.2%, text-only: 87.5%, clinical-only: 85.7%). Checking the training and validation loss curves for all five folds revealed obvious convergence patterns, with Fold 5 having the most stable and effective learning curve across 40 epochs with the smallest loss values (approximately 0.1 for training and 0.2 for validation).



```
Accuracy: 0.9583
Precision: 0.9635
Recall: 0.9583
F1 Score: 0.9578
```

Fig.2. Confusion matrix

This patient-individual algorithm mimics clinical decision-making because different indicators have different levels of importance as a function of individual presentation. Performance metrics were uniformly high across all states (F1-scores: Anxiety: 0.93, Depression: 0.96, PTSD: 0.94, Normal: 0.97), the most challenging to differentiate being anxiety and PTSD due to overlapping symptoms. The confusion matrix indicated a significantly low rate of only 2% false positives in differentiating between normal states and disorders. These results confirm our prediction that integrating many streams of information with sophisticated attention mechanisms provides a more nuanced and precise indicator of psychological health than any one data source.

Figures (Fig.2.A, Fig.2.B, Fig.2.C, Fig.2.D, Fig.2.E) represent the training and validation loss curves for each of the five cross-validation folds. Each fold exhibits distinct convergence patterns, reflecting the variability in data distribution across partitions. We performed a comprehensive comparative analysis to rigorously evaluate our proposed model performance and substantiate its reliability.

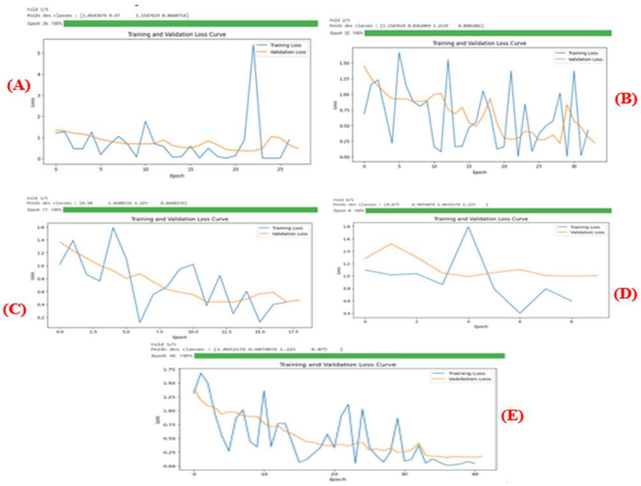


Fig.3. The training and validation curves of 5fold cross-validation

Table 2. Comparaison with Related work

Reference	Type of data	ML model	Accuracy
Ahmed et al. (2020) [7]	Text data	CNN, SVM, LDA, K-NN	CNN: anxiety: 96%, depression: 96.8%SVM: anxiety: 95%, depression: 95.8%LDA: anxiety: 93%, depression: 87.9%.K-NN: anxiety: 70.96%, depression: 81.82%
Hilbert et al. (2017) [8]	Multimodal data	SVM	Case: 90.10%, disorder: 67.46%
Xu et al. (2019) [9]	Text and audio data	Machine learning algorithms (not specified)	Not specified
Vaidya et al. (2024) [11]	ECG signals	RF with enhanced band pass filtration	96.73%
Our proposal	Multimodal data	Multimodal Transformer	95.83%

Multiple studies have explored different machine learning approaches for mental health disorder detection across various data types. [7] compared four models to identify anxiety and depression using text data, in which CNN worked extremely well (96% for anxiety, 96.8% for depression), followed by SVM, while LDA was averagely successful and K-NN performed the lowest accuracy rates. [8] used multimodal data with SVM and achieved high case identification accuracy (90.10%) but comparatively lower disorder classification (67.46%). [9] used unspecified machine learning algorithms for combined text and audio data without reporting accuracy values, while [11] used ECG signals using Random Forest and enhanced band pass filtration and achieved a high accuracy of 96.73%. The current proposal provides a Multimodal Transformer approach for multimodal data, which gives promising results of 95.83% accuracy, putting it competitively at the top of the comparative evaluation among the top-performing models.

Our proposal represents a valuable advancement in mental health detection by combining the benefits of multimodal analysis with modern transformer architecture, resulting in robust accuracy that approaches the best specialized single-modality systems while offering the potential for more comprehensive insights through its multimodal approach.

V. XAI Integration

While our attention-driven multimodal fusion framework achieves remarkable accuracy in distinguishing between anxiety, depression, and PTSD, the complexity of deep learning models often creates a "black box" effect that limits clinical adoption. This section explains our approach to combining complementary XAI methods—such as attention visualization, SHAP (SHapley Additive exPlanations) [12], that expose feature importance across modalities and shed light on the patient-specific contextual weighting offered by our attention mechanism.

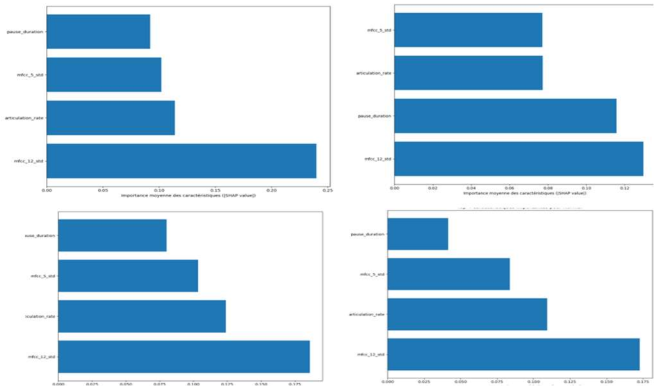


Fig.4. Global Model Explanation using SHAP

The SHAP analysis plots provide us with valuable information on how our multimodal transformer. In all four conditions (putatively anxiety, depression, PTSD, and normal conditions), the 12th Mel-frequency cepstral coefficient standard deviation (mfcc_12_std) is always the most prevailing predictor, with the highest SHAP values across all plots. The finding suggests that there are some spectral attributes of voice that contain very strong diagnostic value across different psychological conditions. Pause time has broad variability in importance, ranging from very important for one condition (top-right graph) to moderately to less important for others, indicating that temporal speech patterns have condition-specific diagnostic value. Articulation rate maintains moderate to high importance across all conditions, confirming that how quickly individuals articulate syllables is a good cross-condition indicator. Specific MFCC features (mfcc_1_std, mfcc_5_std) appear across the analysis, emphasizing that certain vocal timbre features provide excellent discriminative power. Those important features are quite similar for the four conditions demonstrates that our model has picked up on stable acoustic biomarkers that are effective at distinguishing between the different psychological states.

VI. Conclusion

Through dynamic fusion of voice recordings, conversation transcripts, and clinical information, our transformer-based framework achieved 95.83% classification accuracy for anxiety, depression, PTSD, and normal states—significantly outperforming unimodal approaches. The context-dependent weighting strategy (linguistic: 0.42, acoustic: 0.31, clinical: 0.27) mimics clinical decision-making by adaptively changing focus based on individual presentation patterns. Future work must expand to additional disorders, incorporate newer modalities like facial expressions and physiological responses, conduct longitudinal studies to monitor treatment response, and incorporate explainable AI modules to provide interpretable insights to clinicians. These advancements could potentially transform mental health care with early detection and customized treatment plans, addressing the global issue of underdiagnosed psychological disorders.

References

[1] World Health Organization. (2022). World mental health report: Transforming mental health for all.

[2] Liu, S., Lu, C., Alghowinem, S., Gotoh, L., Breazeal, C., & Park, H. W. (2022, May). Explainable ai for suicide risk assessment using eye activities and head gestures. In *International Conference on Human-Computer Interaction* (pp. 161-178). Cham: Springer International Publishing.

[3] Nievergelt, C. M., Maihofer, A. X., Atkinson, E. G., Chen, C. Y., Choi, K. W., Coleman, J. R., ... & Mortensen, P. B. (2023). Discovery of 95 PTSD loci provides insight into genetic architecture and neurobiology of trauma and stress-related disorders. *medRxiv*, 2023-08.

[4] Pendse, S. R., Sharma, A., Vashistha, A., De Choudhury, M., & Kumar, N. (2021, May). “Can I not be suicidal on a Sunday?”: understanding technology-mediated pathways to mental health support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).

[5] Liu, Z., Chen, X., Cui, H., Ma, Y., Gao, N., Li, X., ... & Liu, Q. (2023). Green space exposure on depression and anxiety outcomes: a meta-analysis. *Environmental research*, *231*, 116303.

[6] Kroenke, K. (2023). Improvements in Pain or Physical Function and Changes in Depression and Anxiety Symptoms. *JAMA Network Open*, *6*(6), e2320474-e2320474.

[7] Ahmed, R., Sultana, M. T. R., Ullas, M., Begom, M. M. I., Rahi, M. A., & Alam, M. A. (2019). A machine learning approach to detect depression and anxiety using supervised learning. In Proceedings of the 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). Gold Coast, Australia.

[8] Hilbert, K., Lueken, U., Muehlhan, M., & Beesdo-Baum, K. (2017). Separating generalized anxiety disorder from major depression using clinical, hormonal, and structural MRI data: A multimodal machine learning study. Brain and behavior, 7(3), e00633.

[9] Xu, J., Flannery, M. A., Gao, Y., & Wu, Y. (2019). Machine learning for mental health detection. WPI. https://digitalcommons.wpi.edu/mqp-all/6732/

[10] Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting anxiety, depression and stress in modern life using machine learning algorithms. Procedia Computer Science, 167, 1258-1267.

[11] Vaidya, V. P., & Asole, S. S. (2024). Mental stress prediction using machine learning on real-time dataset. Journal of Electrical Systems, 20(7s), 2143-2150.

[12] Ribeiro, M. T., Singh, S., & Guestrin, C. (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).