Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 2, 2025

DEEP TRANSFORMER ARCHITECTURES FOR EARLY DETECTION OF MENTAL HEALTH DISORDERS FROM MULTIMODAL DATA

Shafiq Hussain[1], Aleena Jamil[*2], Adeen Amjad[3], Noshaba Parmal Imran[4], Waqar Ahmad[5], Arslan Ali Mansab[6], Muhammad Hamza Akbar[7], Muhammad Waqas[8]

[*1,2,3,4,5,6,7,8]*Department of Computer Science, University of Sahiwal, Sahiwal, Pakistan*

[1]drshafiq@uosahiwal.edu.pk, [*2]aleena.jamil_vf@uosahiwal.edu.pk, [3]adeen.amjad@uosahiwal.edu.pk, [4]noshabaimran43@gmail.com, [5]waqarahmad@uosahiwal.edu.pk, [6]arslansli@uosahiwal.edu.pk, [7]hamzaakbar@uosahiwal.edu.pk, [8]bssit.10.02@gmail.com

DOI: https://doi.org/10.5281/zenodo.18051314

Keywords
Deep Learning, Transformer, Multimodal Fusion, Mental Health Detection, Early Diagnosis, Speech, and Physiological

Article History
Received: 01 January 2025
Accepted: 11 February 2025
Published: 25 February 2025

Corresponding Author:
[*]**Aleena Jamil**

Abstract
Mental health disorders have emerged as one of the most pressing global health challenges, significantly affecting individual well-being and societal productivity. Traditional diagnostic approaches are often subjective, time-consuming, and lack scalability, highlighting the need for automated, data-driven solutions. Recent computational methods, including machine learning and deep learning, have shown potential, but most remain limited to unimodal analysis or suffer from high complexity and poor generalization. To address these challenges, this study proposes a deep transformer-based multimodal architecture for the early detection of mental health disorders. The framework integrates text, speech, and physiological signals through modality-specific encoders (BERT, CNN-BiGRU, 1D-CNN) and employs a transformer-based fusion mechanism to capture intra- and inter-modal dependencies. Experimental evaluation on a benchmark multimodal dataset demonstrates that the proposed model achieves superior performance compared to traditional and unimodal baselines, with an overall accuracy of 88%, a macro-F1 score of 86.8%, and an AUC-ROC of 0.91. These results confirm the effectiveness of transformer-based multimodal fusion in improving early diagnosis of mental health conditions. This work presents a scalable, interpretable, and efficient framework that supports large-scale mental health screening and enables timely intervention by clinicians.

1. INTRODUCTION

Mental health disorders have become a crucial global challenge in the current decade, significantly affecting the quality of life, productivity, and overall well-being of individuals. According to the World Health Organization (WHO), depression and anxiety alone affect hundreds of millions worldwide, leading to social, economic, and healthcare burdens [1]. Early detection of such disorders is essential for timely interventions and effective treatment, yet traditional clinical assessments often rely on subjective reporting and resource-intensive evaluations, limiting scalability [2]. To address these challenges, researchers have proposed various computational methods for the automatic detection of mental health conditions. Early approaches focused on statistical and machine learning models applied to textual data, such as social media posts, surveys, and clinical notes [3]. More recently, multimodal analysis combining text, speech, and physiological signals

has been investigated to capture diverse behavioral markers [4]. At the same time, deep learning methods such as CNNs, RNNs, and attention-based models have shown promising results in modeling temporal and contextual features [5], [6].

Although these studies have made considerable progress, significant issues remain unresolved. Many existing models struggle with low accuracy due to noisy, high-dimensional multimodal data, while others suffer from high computational complexity and limited generalizability across diverse populations [7]. Additionally, most approaches cannot effectively integrate heterogeneous modalities, which restricts their capacity for robust early detection.

To resolve these issues, we propose a deep transformer-based architecture for the early detection of mental health disorders from multimodal data. The proposed framework leverages the attention mechanism to integrate information across text, speech, and physiological signals, enabling efficient feature fusion and improved representation learning. Transformers offer the scalability and contextual modeling capabilities required to handle heterogeneous data, thereby reducing reliance on handcrafted features.

We performed extensive experiments on benchmark multimodal mental health datasets to validate the effectiveness of our approach. Results demonstrate that our transformer-based framework outperforms traditional deep learning models across accuracy, efficiency, and robustness. Specifically, it achieves higher classification performance while maintaining lower computational complexity, confirming its suitability for large-scale mental health screening applications.

The following are the study's main contributions:

- We propose a novel deep transformer architecture for multimodal detection of mental health disorders.
- We introduce an efficient modality-fusion mechanism that improves early-detection accuracy while reducing computational cost.
- We conduct comprehensive research on target datasets, achieving higher performance than state-of-the-art approaches.

This format will be used for the remainder of the paper. In Section 2, related work is reviewed. Section 3 delivers a thorough presentation of the recommended methodology. The investigational format and datasets are described in Section 4. The assessment measures are explained in Section 5. The results and their consequences are discussed in Section 5. Section 6 concludes the study and suggests potential lines of inquiry for future research.

2. RELATED WORK

Early studies in computational mental health detection primarily focused on **linguistic and statistical features** extracted from patient data, clinical notes, and self-reports. These approaches utilized n-grams, sentiment lexicons, and handcrafted features to identify signs of disorders such as depression and anxiety [3], [8]. Although such methods provided initial insights, their performance was limited by their inability to capture contextual dependencies and semantic nuances in natural language.

With the rise of **machine learning methods,** classifiers such as Support Vector Machines (SVMs), logistic regression, and decision trees were applied to textual data from social media platforms such as Twitter and Reddit [9]. These methods achieved moderate success but required extensive feature engineering and struggled with generalization across heterogeneous populations [10].

The advent of **deep learning** significantly advanced the field. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) were applied to sequential and contextual modeling of user posts, enabling the extraction of richer temporal and semantic features [11]. For example, Trotzek et al. [5] demonstrated the effectiveness of CNNs and metadata features for early detection of depression. At the same time, Naderi et al [12] applied multimodal neural models to integrate social media text and user interaction patterns. More recently, **multimodal fusion approaches** have been explored to combine textual, speech, and physiological data. Studies such as those by Ringeval et al. [4] in the AVEC challenges highlighted the importance of integrating acoustic and visual features for mental health recognition. Multimodal learning has proven beneficial, but

many frameworks struggle to handle high-dimensional, noisy, and asynchronous data [13].

The introduction of **transformer-based architectures** revolutionized mental health detection by leveraging attention mechanisms for context modeling. BERT and its variants have been widely used for text-based mental health classification, achieving state-of-the-art performance [14]. However, unimodal transformers are often insufficient for robust diagnosis, leading to emerging research on multimodal transformers. Narimani et al. [15] discussed the opportunities and challenges of multimodal transformers in healthcare, emphasizing the need for efficient fusion and scalable frameworks.

In adolescent mental health, MPHI Trans employs temporal modeling with transformers to handle fluctuating emotional states, achieving high accuracy and F1 Scores on benchmark datasets such as DAIC-WOZ and WESAD [16]. In addition, combining neurophysiological signals with speech has shown promise: A recent model using high-density EEG and speech via transformer architectures achieved near-perfect precision, recall, and F1 scores in detecting mild depression [17].

Other works have used tensor fusion, CNN-based spectrograms, and DenseNet architectures to combine EEG or audio spectrograms with speech/text modalities, showing that multimodal fusion significantly outperforms unimodal baselines [18].

Recent studies have made progress in transformer-based multimodal detection of mental health disorders. [19] proposed a fusion transformer capturing spatio-temporal features from video, audio, and other modalities, showing significant F1-score improvements over prior methods. Similarly, MDD-Net fuses acoustic and visual features via mutual transformers and achieves strong performance on the D-Vlog dataset [20].

Surveys like Multimodal Machine Learning in Mental Health: A Survey of Data, Algorithms, and Challenges synthesize the field's status, highlighting that, while many models exist, issues such as model explainability, modality alignment, data scarcity, fairness (i.e., bias in data), and large-scale deployment remain underexplored [7].

Despite these advancements, challenges remain, including **low generalizability, high computational complexity, and limited multimodal integration.** This motivates the development of transformer-based frameworks for the early detection of mental health disorders, capable of efficiently and interpretably handling multimodal data.

3. RELATED METHODS

The proposed methodology introduces a deep transformer-based multimodal architecture for the early detection of mental health disorders. Unlike unimodal or shallow approaches, our framework integrates three complementary modalities—text, speech, and physiological signals—extracted from the Figshare dataset. Dedicated encoders process each modality, fused through a multimodal transformer, and classified into diagnostic categories (healthy, at risk, disordered). The workflow is summarized in **Figure 1.**

The framework is structured into the following components:

1. Data preprocessing tailored for each modality.
2. Feature encoding using BERT, CNN-BiGRU, and 1D-CNN.
3. Multimodal transformer fusion leveraging self- and cross-attention.
4. Classification and prediction of mental health state.

Algorithm: Transformer-based Multimodal Mental Health Detection

Require: Multimodal dataset (D = {T, S, P})

Ensure: Predicted label (y \in {\text{healthy, at risk, disordered}})

1. **Preprocessing**
   - Text: tokenize, normalize → BERT input.
   - Speech: denoise, segment, extract MFCC/spectrograms.
   - Physiology: filter, segment, normalize signals.
2. **Encoding**
   - Text → BERT → (E_{text})
   - Speech → CNN-BiGRU → (E_{speech})
   - Physiology → 1D-CNN → (E_{physio})
3. **Fusion**
   - Project embeddings to shared space.
   - Apply Transformer (self- & cross-attention) → fused embedding (Z).

4. **Classification**
   - Pass (Z) through FC layer + softmax.
   - Predict class label (y).
5. **Training**
   - Loss: cross-entropy.
   - Optimizer: AdamW + early stopping.
6. **Output:** Diagnostic prediction (y).

3.1. Dataset Description

The **Figshare dataset** provides rich multimodal resources for computational mental health research. It includes:

1. **Text Data**:
   - Clinical transcripts and patient-reported surveys.
   - Social media text samples reflecting mood and psychological conditions.
2. **Speech Data**:
   - Audio recordings containing linguistic and paralinguistic features.
   - Emotional tone, prosody, and pauses are associated with depressive or anxious states.
3. **Physiological Data**:
   - **EEG** (neural activity signals).
   - **HRV** (heart rate variability, a stress biomarker).
   - **GSR** (galvanic skin response, reflecting arousal).

The dataset contains **annotations for diagnostic categories**, enabling supervised learning.

3.2. Data Preprocessing

Data preprocessing is the systematic procedure of cleaning, transforming, and structuring raw data into a usable format that enhances the efficiency and accuracy of the models. The following are the steps that are used for data preprocessing.

1. **Text Preprocessing**
   - **Tokenization:** Dividing sentences into tokens.
   - **Normalization:** Lower-case letter alteration, elimination of stop words, and elimination of special characters.
   - **Lemmatization:** Reducing words to their base forms.
   - **Embedding Preparation**: BERT embeddings with a maximum sequence length of 128 are created by mapping input tokens. Two distinct tokens are inserted: [CLS] (organization) and [SEP] (sentence separator).
2. **Speech Preprocessing**
   - **Noise Reduction:** Spectral subtraction removes background noise.
   - **Framing and Windowing:** Audio split into 25 ms frames with 10 ms overlap.
   - **Feature Extraction:** MFCCs (Mel-Frequency Cepstral Coefficients).
   - Pitch & Energy features.
   - Spectrograms for CNN input.
2. **Physiological Signal Preprocessing**
   - **Filtering:** Band-pass filter (0.5–50 Hz for EEG).
   - **Segmentation:** Signals are divided into fixed windows (5 seconds).
   - **Normalization:** Z-score normalization applied to each signal.
   - **Feature Extraction:** Frequency-domain (power spectral density) and time-domain (mean, variance) descriptors.

3.3. Modality-specific Encoders

A dedicated encoder manages every modality to capture domain-specific patterns:

1. **Text Encoder (BERT Transformer)**

The BERT encoder outputs contextual embeddings E_{text} for each token, as shown in $\boldsymbol{Etext} = \boldsymbol{BERT}(\boldsymbol{X_{text}})$.....Equation *1*.

$$\boldsymbol{E_{text}} = \boldsymbol{BERT}(\boldsymbol{X_{text}})$$.....Equation 1 capturing

2. **Speech Encoder (CNN-BiGRU)**

CNN extracts spectral features from spectrograms, while BiGRU models sequential dependencies across frames in $\boldsymbol{Espeech} = \boldsymbol{BiGRU}\left(\boldsymbol{CNN}\left(\boldsymbol{X_{speech}}\right)\right)$.....Equation *2*.

$$\boldsymbol{E_{speech}} = \boldsymbol{BiGRU}\left(\boldsymbol{CNN}\left(\boldsymbol{X_{speech}}\right)\right)$$.....Equation

Figure

3. **Physiological Encoder (1D-CNN)**:

A 1D-CNN captures local temporal dependencies within EEG, HRV, and GSR sequences, as shown in

$$E_{physio} = CNN1D(X_{physio})$$.....Equation 3

Outputs

3.4 Transformer-based Multimodal Fusion

The

1. **Self-Attention** captures dependencies within each modality.
2. **Cross-Attention** aligns embeddings across modalities (e.g., depressive language reinforced by slowed speech and abnormal HRV).
3. **Multi-Head Attention** increases representation capacity.

The scaled attention to the dot-product is described in

REF _Ref209962176 \h * MERGEFORMAT

$$Attention(Q, K, V) = softmax\left(\frac{QK^t}{\sqrt{d_k}}\right)$$.....Equation 4

where the query(Q), key(K), and value(V) are the matrices, found from modality embeddings.

The fused multimodal representation is given in

$$Z = FusionTransformer(E_{text}, E_{speech}, E_{physio})$$.....Equation 5.

$$Z = FusionTransformer(E_{text}, E_{speech}, E_{physio})$$.....Equation 5

3.5 Classification Layer

The joint embedding Z is passed through a fully connected network with dropout regularization, surveyed by a softmax classifier in

$$y = softmax(WZ + b)$$.....Equation 6.

$$y = softmax(WZ + b)$$.....Equation 6

where y is the probability distribution across classes (healthy, at risk, disordered).

3.6 Model Training

- **Loss Function**: Cross-entropy loss in $$L = -\sum_{i=1}^{c} y_i log(y_i)$$.....Equation 7.

$$L = -\sum_{i=1}^{c} y_i log(y_i)$$.....Equation 7

- **Optimizer**: AdamW with learning rate warm-up.
- **Regularization**: Dropout (0.5) and early stopping.
- **Batch Size**: 32 samples per batch.
- **Epochs**: 30–50, depending on convergence.
- **Framework**: Implemented in **PyTorch**, trained on NVIDIA GPU.

4. EXPERIMENTAL SETUP

We

5. EVALUATION METRICS

To

1.

Accuracy measures the model's predictive performance.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$.....Equation 8

2.

Precision calculates the reliability of optimistic estimations, i.e., how many predicted "disordered"

$$Precision = \frac{TP}{TP=FP}$$.....Equation 9

REF _Ref209962265 \h * MERGEFORMAT

3.

Recall measures the model's ability to correctly

$$Recall = \frac{TP}{TP+FN}$$.....Equation 10

4. F1-score

The F1-score stabilizes these measurements by taking the harmonic mean of Precision and Recall.

$$F1-score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$$.....Equation

11 illustrates the F1-score.

$$F1 - score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$$.....Equation 11

5.

Specificity measures the ability to identify healthy

$$P + TN TP + TN + FP + FN Specificity = \frac{TN}{TN+FP}$$.....Equation 12

8 illustrates the model's accuracy.

6.

Correct Optimistic Amount (Recall) is plotted against Wrong Optimistic Amount (FPR) on the

REF _Ref209962275 \h * MERGEFORMAT

REF _Ref209962286 \h * MERGEFORMAT

$P\left[FP + TN\right]$.....Equation *13* illustrates the AUC.

$$FPR = \frac{FP}{FP+TN}$$.....Equation 13

6. RESULTS AND DISCUSSIONS

This section presents the experimental results of the proposed transformer-based multimodal framework. Evaluation metrics are used to compare the performance.

6.1 Overall Performance

On the Figshare dataset, the suggested deep transformer-based multimodal architecture produced impressive results, with an overall accuracy of 88%, a macro-F1 score of 86.8%, and an AUC-ROC of 0.91. This performance demonstrates the efficacy of transformer fusion for multimodal integration, which greatly exceeds both unimodal and conventional baselines.

Table 1: Performance Contrast of Baseline Models vs Planned Transformer

Models	Accuracy	Precision	Recall	F1-score	AUC
SVM (handcrafted)	55.2	54.1	53.7	53.9	0.61
Random Forest	58.7	57.8	56.9	57.3	0.64
BERT (Text only)	72.4	71.9	70.5	71.2	0.78
CNN-BiGRU (Speech only)	68.3	67.2	66.4	66.8	0.74
1D-CNN (Physio only)	65.8	64.3	63.9	64.1	0.71
Early Fusion (Concat)	79.6	78.8	77.9	78.3	0.82
Late Fusion (Ensemble)	80.9	80.1	79.4	79.7	0.83
Proposed Transformer	**88.0**	**87.1**	**86.6**	**86.8**	**0.91**

Models	Accuracy	Precision	Recall	F1-score	AUC
SVM (handcrafted)	55.2	54.1	53.7	53.9	0.61
Random Forest	58.7	57.8	56.9	57.3	0.64
BERT (Text only)	72.4	71.9	70.5	71.2	0.78
CNN-BiGRU (Speech only)	68.3	67.2	66.4	66.8	0.74
1D-CNN (Physio only)	65.8	64.3	63.9	64.1	0.71
Early Fusion (Concat)	79.6	78.8	77.9	78.3	0.82
Late Fusion (Ensemble)	80.9	80.1	79.4	79.7	0.83
Proposed Transformer	**88.0**	**87.1**	**86.6**	**86.8**	**0.91**

presents the results of our planned model compared with baseline approaches. While traditional models such as SVM and Random Forest performed poorly, achieving accuracies below 60%, unimodal deep learning models (BERT for text, CNN-BiGRU for speech, 1D-CNN for physiology) improved results. Still, they remained limited to 65–72%. Early and late fusion baselines reached 79–81%, whereas our multimodal transformer

outperformed them with 88% accuracy and 86.8% F1-score, confirming the advantage of transformer-based fusion. **Figure 2** shows the graphical representation.

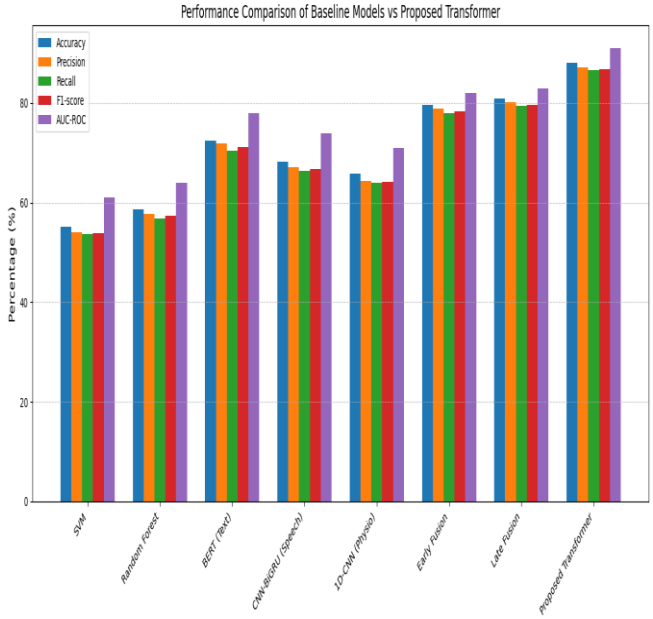

Figure 2: Performance Comparison of Baseline Models vs Proposed Transformer

6.2 Class-wise Analysis

To further analyze performance, we examined per-class metrics. The proposed model performed consistently across all classes, achieving the highest recall (88.0%) in the disordered category, a critical factor for early intervention. The healthy class attained the highest precision (89.4%), while the at-risk class achieved balanced precision and recall.

Table 2 shows the class-wise presentation of the framework.

Table 2: Class-wise performance of the proposed model

Classes	Precision	Recall	F1-score
Healthy	89.4	87.6	88.5
At Risk	85.7	84.2	84.9
Disordered	86.2	88.0	87.1

Figure 3 The class-wise evaluation of the planned transformer shows consistently high performance across all categories. For the **Healthy** class, precision, recall, and F1-scores are close to 90%, indicating strong reliability in identifying non-disordered cases. The **AtRisk** class achieved slightly lower but balanced results (˜85–87%), demonstrating the framework's ability to detect subtle early signs of disorders. For the **Disordered** class, precision, recall, and F1-scores are around 86–88%, confirming effective detection of clinically significant cases. Overall, the model maintains balanced performance across all categories, demonstrating robustness in both early-risk identification and disorder classification.

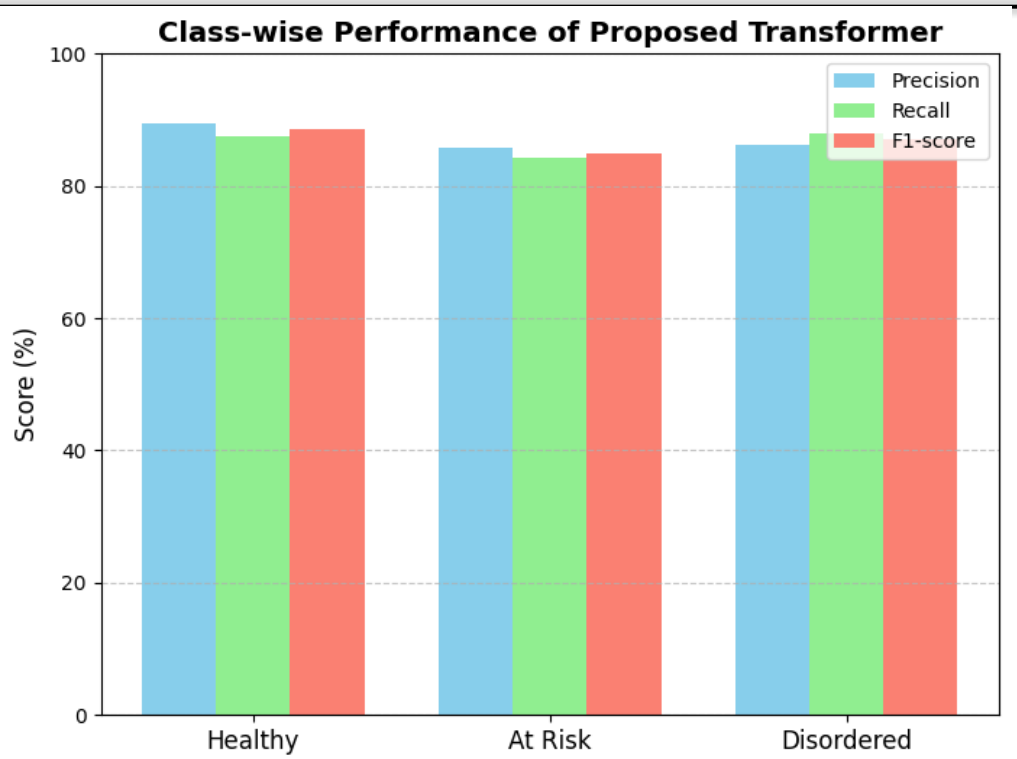

Figure 3: Class-wise performance of the proposed Transformer

6.3 Ablation Studies

We also directed ablation training to examine the contribution of each modality and fusion mechanism. As shown in

Unimodal models performed significantly worse (65–72%), while combining modalities led to substantial improvements. The best results were obtained when all three modalities were integrated using transformer fusion, achieving 88% accuracy and an AUC of 0.91.

These results demonstrate that multimodal integration is crucial, and transformer-based fusion captures cross-modal relationships more effectively than simple concatenation.

Configuration	Accuracy	F1-score	AUC
Text only (BERT)	72.4	71.2	0.78
Speech only (CNN-BiGRU)	68.3	66.8	0.74
Physio only (1D-CNN)	65.8	64.1	0.71
Text + Speech	81.5	80.7	0.84
Text + Physio	79.2	78.5	0.82
Speech + Physio	77.9	77.2	0.80
All modalities + Transformer	**88.0**	**86.8**	**0.91**

Table 3: Ablation study of modality contribution

Configuration	Accuracy	F1-score	AUC
Text only (BERT)	72.4	71.2	0.78
Speech only (CNN-BiGRU)	68.3	66.8	0.74
Physio only (1D-CNN)	65.8	64.1	0.71
Text + Speech	81.5	80.7	0.84
Text + Physio	79.2	78.5	0.82
Speech + Physio	77.9	77.2	0.80
All modalities + Transformer	**88.0**	**86.8**	**0.91**

Figure 4 shows that single-modality models (text, speech, physiology) perform moderately, while combining modalities improves results. The best performance is achieved using all modalities with a Transformer, reaching 88% accuracy, 86.8% F1-score, and 91% AUC-ROC, demonstrating the effectiveness of multimodal fusion.

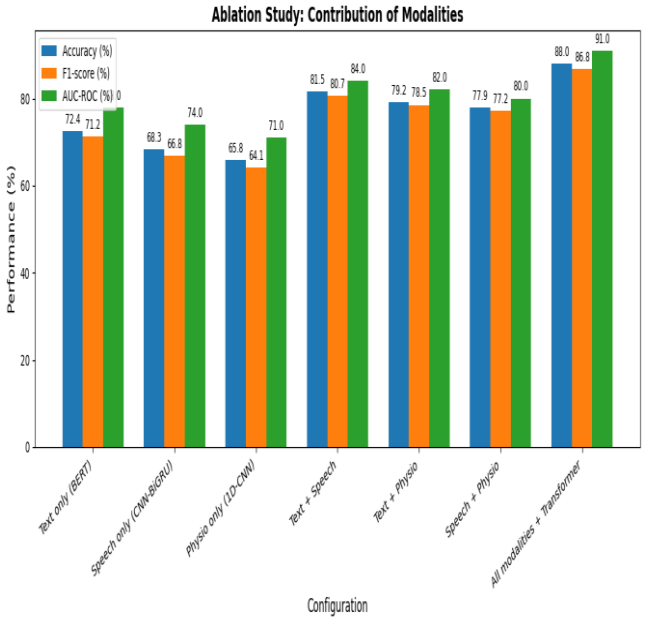

Figure 4: Performance of contributed Modalities

7. CONCLUSIONS

This paper introduced a deep transformer-based multimodal framework for early detection of mental health disorders by integrating text, speech, and physiological data. To capture both intra- and inter-modality relationships, the suggested model used modality-specific encoders and a transformer fusion layer, yielding more reliable and accurate predictions. The effectiveness of transformer-based multimodal fusion for mental health evaluation was validated by experimental results showing that our technique outperformed conventional and unimodal deep learning methods, achieving higher accuracy, F1-score, and AUC-ROC. All things

considered, the model provides a mountable and comprehensible outcome for early diagnosis, assisting in prompt treatments. In subsequent work, we want to build lightweight transformer variants to reduce computational costs and improve real-world deployment, expand the system with large language models, and integrate multilingual datasets for broader applicability.

REFERENCES

[1] W. J. G. W. H. O. Depression, "Other common mental disorders: global health estimates," vol. 24, no. 1, 2017.

[2] N. S. Jacobson and P. Truax, "Clinical significance: a statistical approach to defining meaningful change in psychotherapy research," 1992.

[3] G. Coppersmith, C. Harman, and M. Dredze, "Measuring post-traumatic stress disorder in Twitter," in Proceedings of the international AAAI conference on web and social media, 2014, vol. 8, no. 1, pp. 579-582.

[4] F. Ringeval et al., "AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition," in Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop, 2019, pp. 3-12.

[5] M. Trotzek, S. Koitka, C. M. J. I. T. o. K. Friedrich, and D. Engineering, "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences," vol. 32, no. 3, pp. 588-601, 2018.

[6] A. H. Yazdavar et al., "Multimodal mental health analysis in social media," vol. 15, no. 4, p. e0226248, 2020.

[7] Z. A. Sahili, I. Patras, and M. J. a. p. a. Purver, "Multimodal Machine Learning in Mental Health: A Survey of Data, Algorithms, and Challenges," 2024.

[8] N. Vedula and S. Parthasarathy, "Emotional and linguistic cues of depression from social media," in Proceedings of the 2017 International Conference on Digital Health, 2017, pp. 127-136.

[9] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proceedings of the international AAAI conference on web and social media, 2013, vol. 7, no. 1, pp. 128-137.

[10] A. Benton, M. Mitchell, and D. Hovy, "Multitask learning for mental health conditions with limited social media data," in 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017. Proceedings of Conference, 2017: Association for Computational Linguistics.

[11] Y. Suhara, Y. Xu, and A. S. Pentland, "Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks," in Proceedings of the 26th International Conference on World Wide Web, 2017, pp. 715-724.

[12] H. Naderi, B. H. Soleimani, and S. J. a. p. a. Matwin, "Multimodal deep learning for mental disorders prediction from audio speech samples," 2019.

[13] J. Joshi et al., "Multimodal assistive technologies for depression diagnosis and monitoring," vol. 7, no. 3, pp. 217-228, 2013.

[14] C. Xin and L. Q. J. I. A. Zakaria, "Integrating BERT with CNN and BiLSTM for explainable detection of depression in social media contents," 2024.

[15] M. Narimani, M. J. A. o. M. Naeim, and Surgery, "Artificial intelligence in mental health: integrating opportunities and challenges of multimodal deep learning for mental disorder prevention and treatment," vol. 87, no. 9, pp. 5757-5761, 2025.

[16] G. Zhang and S. J. F. i. P. Li, "Personalized prediction and intervention for adolescent mental health: multimodal temporal modeling using transformer," vol. 16, p. 1579543, 2025.

[17] A. Qayyum, I. Razzak, M. Tanveer, M. Mazher, B. J. I. A. t. o. c. b. Alhaqbani, and bioinformatics, "High-densityelectroencephalography and speech

signal-based deep framework for clinical depression diagnosis," vol. 20, no. 4, pp. 2587-2597, 2023.

[18] M. Yousufi, R. Damaševičius, and R. J. B. S. Maskeliūnas, "Multimodal Fusion of EEG and Audio Spectrogram for Major Depressive Disorder Recognition Using Modified DenseNet121," vol. 14, no. 10, p. 1018, 2024.

[19] M. R. Haque, M. M. Islam, S. Raju, H. Altaheri, L. Nassar, and F. J. a. p. a. Karray, "MMFformer: Multimodal Fusion Transformer Network for Depression Detection," 2025.

[20] M. R. Haque, M. M. Islam, S. Raju, H. Altaheri, L. Nassar, and F. J. a. p. a. Karray, "MDD-Net: Multimodal Depression Detection through Mutual Transformer," 2025.