SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

CM-CAT: A Cross-Modal Contextual Attention Transformer for Robust Mental Health Prediction via Textual and Facial Temporal Dynamics

Ms. Priyanka P. Boraste[1], Dr. Monika S Deshmukh[2]

[1]Research Scholar, School of computer Science and Engineering, Sandip University, Nashik 422213, Maharashtra, India. Orcid: 0009-0008-2565-1455, E-mail: boraste.priyanka@kbtcoe.org

[2]Research Guide, School of computer Science and Engineering, Sandip University, Nashik 422213, Maharashtra, India. Orcid: 0009-0009-6185-8768, E-mail: monika.deshmukh@sandipuniversity.edu.in

Abstract

The mental health crisis due to the COVID-19 pandemic has been getting worse globally, and there is an urgent need for new easy-to-use diagnostic techniques for objective depression screening. Deep learning has been used to automate the process of identifying emotions, but the sensitivity and specificity needed to capture the important micro-expressions are not often addressed. Most models fail to capture the nuanced micro-dynamics that can differentiate true depressed affect from social masking. This paper presents Cross-Modal Contextual Attention Transformer (CM-CAT) for enhanced mental health prediction that integrates sentiment in text and facial expression temporal dynamics. The architecture is based on a dual-stream long-range cross attention (visual and linguistic) capture model. The proposed transformer architecture model is used with DepVidMood for validation. CM-CAT model is further enhanced with an Explainable AI (XAI) which allows non-expert users to gain insight into the model and its prediction. The CM-CAT model framework achieved the highest accuracy of 92.4% and F1-score of 0.89 across 4 severity classes of mental health (Healthy, Mild, Moderate and Severe). The CM-CAT model was also able to surface and focus temporal saliency around the clinically significant periorbital and perioral areas salient consistent with established biomarkers. This study enhanced CM-CAT model has a clinically interpretable design that supports the earliest intervention for mental health concerns and enables objective psychiatric evaluations.

Objectives: The expanding global mental health burden has intensified the critical need for objective, scalable diagnostic screening tools for clinical depression. While conventional deep learning models can automate basic affect recognition, they frequently fail to capture the highly subtle facial micro-dynamics that differentiate genuinely depressed affect from superficial social masking. This study presents a novel, robust multimodal framework designed to solve cross-modal temporal dependencies and enhance prediction reliability.

Methods: We present the Cross-Modal Contextual Attention Transformer (CM-CAT), which integrates contextual text semantics with detailed facial temporal dynamics through a distinct dual-stream architecture. The pipeline merges a spatio-temporal visual stream built on facial Action Units and a Temporal Convolutional Network with a contextual textual stream utilizing RoBERTa embeddings. Evaluation was executed on the DepVidMood clinical video corpus across four severity tiers, utilizing a person-independent 70%/15%/15% train/validation/test split.

Results: On the held-out test partition, CM-CAT achieves a high 92.4% validation accuracy and a macro-averaged F1-score of 0.89, outperforming contemporary multimodal baseline models. Furthermore, an Explainable AI (XAI) module based on integrated gradients demonstrates that attention accurately concentrates on the periorbital and perioral zones associated with established physiological markers of depressive affect.

Conclusions: CM-CAT introduces a highly interpretable, reliable, and clinically viable digital biomarker tool that successfully advances objective mental health tracking methodologies inside resource-constrained psychiatric informatics networks.

Keywords: Affective computing; Cross-modal fusion; Transformer; Depression detection; Explainable AI (XAI); Facial action units.

1. Introduction

Major Depressive Disorder (MDD) is one of the most serious mental-health conditions worldwide. It is a multifaceted psychiatric disorder that affects brain regions governing thought, emotion, and behavioural expression. The rising incidence of depression and anxiety places an enormous burden on the economy and the healthcare system, underlining the urgent need for cost-effective, easily deployable, and automated diagnostic tools. Conventional diagnostic frameworks are challenged by the fact that most clinical assessments rely on subjective clinician interviews and self-administered questionnaires such as the Patient Health Questionnaire-9 (PHQ-9) and the Beck Depression Inventory-II (BDI-II). The inconsistency and subjective bias of such assessments limit the scalability of mental-health evaluation, and the need for trained personnel together with the lack of real-time assessment creates a significant barrier to timely psychiatric evaluation [1]. Artificial intelligence (AI) offers a way to address these challenges by enabling real-time assessment while reducing dependence on scarce clinical expertise. With recent advances in machine

learning (ML), there is growing interest in analysing facial, vocal, and textual signals, alongside other physiological variables, to characterise mental-health status. Compared with unimodal approaches, integrating several modalities are expected to provide a more holistic basis for diagnosis and prognosis [2]. Despite this progress, extracting meaningful and discriminative representations remains difficult, particularly for multimodal data spanning video, audio, and text. Because depression is a long-term, temporally evolving condition, models that focus only on static or short-window features may miss clinically important dynamics. The combination of signal processing and natural language processing (NLP) is a promising avenue for identifying emotional and behavioural cues within linguistic and prosodic components. Transformer models, in particular, are well suited to capturing long-range dependencies and feature interactions across modalities [3]. In clinical communication, understanding the patient's emotional state is essential for appropriate diagnosis and treatment. Contextual embeddings and multimodal representations have improved the detection of latent psychological states, especially through transformer-based sentiment analysis. A limitation, however, is that many conventional transformer models rely on pairwise attention, which restricts their ability to integrate interactions across modalities and limits the depth of information fusion [4].

Recent research has therefore developed more sophisticated multimodal transformers with cross-attention mechanisms that capture inter-modal relationships more robustly. Text-guided cross-attention transformers and tensor-based multimodal transformers provide stronger fusion by jointly modelling relationships across dimensions, and networks that combine sequential and spatial attention can more effectively capture temporal flow and cross-modal interdependencies for depression detection [5]. In spite of these advances, interpretability, data scarcity, and practical deployment remain concerns. Explainable AI (XAI) methods are increasingly used so that model inferences can be presented in a human-understandable form. Together with non-invasive multimodal sensing, these directions support the goal of scalable, clinically relevant mental-health monitoring.In this context, we introduce CM-CAT, the Cross-Modal Contextual Attention Transformer, which combines the sequential dynamics of facial expression with text semantics. The framework is designed to be scalable, interpretable, and relevant to clinical practice, using temporal saliency and contextual attention. The main contributions of this work are summarised below:

A text-guided cross-modal contextual attention mechanism in which linguistic context queries the visual stream, allowing the model to prioritise clinically relevant facial micro-expressions and to flag affective incongruence ("social masking").

• A dual-stream design that couples a Temporal Convolutional Network over facial Action Units with RoBERTa-based contextual text embeddings, modelling depression as a temporally evolving process rather than a static snapshot.

• An integrated XAI module that produces spatial saliency maps and attention–Action-Unit correlations, offering clinically interpretable evidence for each prediction.

• An evaluation on the DepVidMood clinical-interview corpus under a person-independent split, together with a transparent benchmarking protocol, hyperparameter reporting, and error analysis to support reproducibility.

The remainder of the paper is organised as follows. Section 2 reviews related work and positions the contribution. Section 3 details the proposed CM-CAT framework. Section 4 describes the experimental setup. Section 5 presents results and discussion, including error and ablation analyses. Section 6 concludes and outlines future work.

2. Related Work

Automated depression detection has evolved from primitive psychiatric assessment to complex multimodal deep-learning frameworks. We review recent developments in machine learning, transformer-based approaches, and explainable AI as they relate to mental-health prediction, and conclude by identifying the gap this work addresses.

2.1From Classical Machine Learning to Deep Learning

Objectivity is a major barrier in clinical evaluation, as many techniques rely on subjective feedback and require highly trained professionals, which makes the process difficult to scale. Early efforts applied classical ML models: Taskynbayeva and Gutoreva [6] reported Random Forest and Gradient Boosting as dominant techniques for anxiety prediction with high accuracy, and Basu et al. [7] combined Support Vector Machines with deep-learning models to analyse verbal and textual data for anxiety patterns. Classical ML, however, still struggles to model the complex non-linear relationships present in psychiatric data. Deep-learning (DL) approaches have shifted the field toward automated, data-driven recognition across data types: Umair et al. [8] demonstrated transformer-based recognition of depression from EEG data, while Huynh and Deshpande [9] noted that DL models in neuroimaging face data-scarcity and population-bias issues.

2.2Developments in Multimodal and Behavioural Analysis

Multimodal systems capture a broader range of behavioural signals than unimodal systems. Zou et al. [10] introduced the Chinese Multimodal Depression Corpus (CMDC), supporting behavioural analysis across visual, auditory, and linguistic components for clinical evaluation. To model depression as a temporal

disorder, Zhao et al. [11] proposed DepressionMIGNN, a graph-learning approach capturing temporal and contextual relationships. Li et al. [12] reported advantages from multimodal fusion of video, audio, and text on clinical datasets, and Li et al. [13] used transformer-based feature fusion for sentiment analysis on social-media data. Modelling contextual and temporal relationships nonetheless remains a challenge.

2.3Transformer Architectures in Affective Computing

Transformer architectures excel at capturing long-range dependencies in multimodal data and have been applied to wearable physiological signals for affective-state recognition, as shown by Li and Zhang [14]. Song et al. [15] developed personalised cross-modal transformers that predict mental-health status from integrated physiological signals. A common limitation, however, is reliance on pairwise attention. Sun et al. [5] addressed this with TensorFormer, which provides tensor-based attention for cross-modal interactions, and Li et al. [13] presented a feature-fusion transformer to align text and image features for sentiment analysis.

2.4 Cross-Modal Attention and Fusion

Emerging studies focus on improving cross-modal attention for depression analysis. Teng et al. [1] introduced a sentiment pre-trained, text-guided multimodal cross-attention transformer that integrates text, audio, and visual components. WavFace, by Flores et al. [4], combines sequential and spatial self-attention for depression screening. Spatio-temporal models have also been explored: Tao et al. [16] proposed DepMSTAT, which quantifies spatial and temporal correlations in multimodal data, and Jia et al. [17] introduced Bidirectional Multimodal Block-Recurrent Transformers (BMBRT) to reduce cross-modality noise. Wei et al. [18] presented CANAMRF, which dynamically adjusts the importance of each modality during fusion, and Yao et al. [19] introduced the Multi-Stage Joint Attention Fusion Network (MSJAFN) for progressive fusion.

II.5 Interpretability and Explainable AI (XAI)

Explainability is increasingly important for clinical adoption. Pandey et al. [20] created an explainable multimodal transformer for text and audio using attribution methods such as integrated gradients, and Kawamura et al. [21] showed that non-contact multimodal emotion recognition can rival clinicians on some tasks. Dialogue-oriented models such as MultiFlipFormer by Sharma et al. [22] capture emotional shifts in therapeutic conversations, and Nithya and Kanna [23] proposed EMOTEX-PR, which uses reinforcement learning and transformer attention for real-time sentiment analysis under accuracy–latency constraints.

II.6 Physiological Signals and Generative AI

Recent work increases modality diversity through physiological signals. Zhu et al. [24] proposed a transformer-based fusion model combining EEG and pupil-area signals for mild-depression recognition. Wang et al. [25] proposed a multimodal transformer graph-convolution attention network; although developed for insomnia-disorder detection, its brain-connectivity feature extraction is methodologically relevant to interpretable multimodal modelling. Umair et al. [8] implemented split learning for decentralised, privacy-preserving EEG-based depression detection, and Ma et al. [26] studied dynamic functional connectivity in chronic insomnia, illustrating connectivity-based analysis of neural dysregulation. Generative AI is also reshaping the field: Mujeeb and Raza [3] integrated generative AI with virtual-reality therapy, and Huynh and Deshpande [9] addressed data scarcity through GAN-based augmentation of neuroimaging. Broader adoption is reflected in multimodal dialogue systems [27], personalised healthcare monitoring [28], and large-language-model-based depression detection [29].

II.7 Research Gap and Positioning

Although progress has been substantial, several gaps remain. Many current models emphasise short-term behavioural trends rather than long-term temporal dynamics, and the explainable fusion of text semantics with high-resolution, temporally dynamic facial expression is still underdeveloped. He M. et al. [30] noted that strong multimodal assimilation across heterogeneous datasets remains difficult. CM-CAT targets this gap by combining text-guided cross-modal contextual attention with temporal-trajectory modelling and built-in interpretability. Table 1 summarises how CM-CAT relates to representative recent models along the capabilities most relevant to clinically interpretable, temporally aware depression assessment.

Table 1: Capability comparison of CM-CAT with representative multimodal depression-detection models.

Model	Temporal modelling	Cross-attention	Explainability	Severity classification
TensorFormer [5]	Limited	Yes (tensor-based)	No	Partial
WavFace [4]	Yes (seq.+spatial)	Partial (self-attn.)	No	Binary
DepMSTAT [16]	Yes (spatio-temporal)	Yes	No	Partial
Text-Guided CAT [1]	Limited	Yes (text-guided)	No	Binary

Model	Temporal modelling	Cross-attention	Explainability	Severity classification
MSJAFN [19]	Partial	Yes (joint attn.)	No	Partial
CM-CAT (proposed)	Yes (TCN)	Yes (text-guided)	Yes (IG + saliency)	Yes (4-class)

III. Proposed Cm-CAT Framework

The CM-CAT framework addresses the difficulty of "social masking" in clinical depression assessment, in which a patient's outward expression may conceal their underlying affective state. Unlike architectures that treat modalities as independent parallel streams, CM-CAT uses a contextualised cross-modal mechanism in which the temporally expressed dynamics of the face are interpreted under the guidance of the sentiment of the accompanying text. This section presents the framework structure and its mathematical formulation.

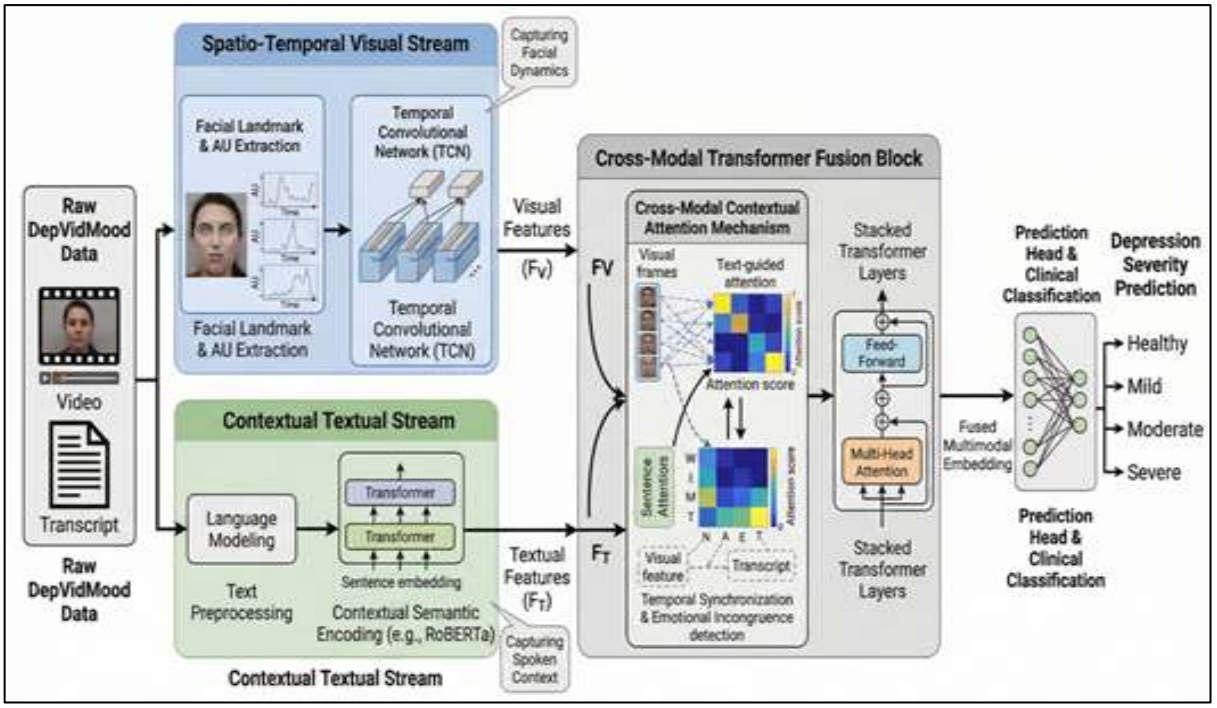

Fig. 1: The architectural pipeline of the CM-CAT framework. Facial images are anonymised to protect participant privacy.

As shown in Figure 1, the pipeline transforms raw input data into a final clinical classification through three components: the spatio-temporal visual stream, the contextual textual stream, and the cross-modal transformer fusion block. The information flow is as follows. The visual stream extracts per-frame facial Action-Unit features and models their temporal evolution with a Temporal Convolutional Network (TCN); the textual stream encodes the interview transcript with RoBERTa; the two streams are projected to a common embedding space; the fusion block applies text-guided cross-modal attention so that linguistic context prioritises clinically relevant visual cues; and the fused representation is passed to a classification head that predicts one of four severity levels.

Visual Stream: Facial Landmark and Action-Unit Extraction

The visual modality is derived from the DepVidMood clinical-interview recordings. To analyse the micro-dynamics of facial expression, we use the MediaPipe Face Mesh API to extract 468 three-dimensional facial landmarks per frame. We focus on the periocular (around the eyes) and perioral (around the mouth) regions, which are clinically associated with flat affect. Let $V = \{f_1, f_2, \ldots, f_t\}$ denote a video of T frames. For each frame f_t, we compute a feature vector x_t^v from Euclidean distances between landmark indices corresponding to Action Units (AUs) such as AU04 (Brow Lowerer) and AU12 (Lip Corner Puller). The Euclidean distance d between two landmark points p and q is:

$$d(p, q) = \sqrt{[\,(p_x - q_x)^2 + (p\_y - q\_y)^2 + (p\_z - q\_z)^2\,]} \quad (1)$$

AU04 and AU12 are standard Facial Action Coding System descriptors, reduced lip-corner pulling (AU12) and increased brow lowering (AU04) are linked to depressive affect in the clinical and affective-science literature, where affiliative AU12 expressions decrease and sadness-related actions such as AU04 are more prominent with greater symptom severity , which motivates their selection here as interpretable visual markers.

Textual Stream: Contextual Embeddings

In parallel, the textual modality L is processed from the interview transcript. We use a pre-trained RoBERTa (Robustly Optimised BERT Approach) model to produce dense semantic embeddings, chosen for its strong performance in capturing the subtleties of clinical conversation. Each token w_i is mapped to a d-dimensional contextual representation:

$h_i^l = \text{RoBERTa}(w_i)$ (2)

Table 2 summarises the input features and their dimensionality for both modalities, including the projection of each stream to a common 512-dimensional space prior to fusion.

Table 2: Summary of input features and dimensionality for CM-CAT

Modality	Feature category	Specific indices / source	Dim. (D)
Visual (spatio-temporal)	Action Unit 04 (Brow Lowerer)	MediaPipe: Dist(P107, P133)	1×T
	Action Unit 12 (Lip Corner Puller)	MediaPipe: Dist(P13, P61)	1×T
	Global facial geometry	468 3D landmark coordinates	1404×T
	Visual sub-total (Dv)	Projected feature vector	512
Textual (contextual)	Token embeddings	RoBERTa-base vocabulary	768×N
	Contextual hidden states	Multi-head self-attention layers	768×N
	Semantic global vector	[CLS] token / pooled output	768×1
	Textual sub-total (Dl)	Projected feature vector	512
Fused modality	Cross-modal attention	Eq. (6) and Eq. (7) interaction	1024

Spatio-Temporal Feature Embedding

To model the time-series trend of a patient's state, the visual features are passed to a Temporal Convolutional Network (TCN). The TCN uses causal convolutions so that the prediction at time t depends only on frames $\{f_1, \ldots, f_t\}$; this prevents temporal leakage and reflects real-time clinical observation. The TCN outputs a sequence of hidden states $H^v = \{h_1^v, h_2^v, \ldots, h_t^v\}$, where each state captures facial dynamics within a local window. To preserve relative temporal position, a sinusoidal positional encoding (PE) is added to both modalities:

$Z^v = H^v + PE$ (3)

$Z^l = H^l + PE$ (4)

Cross-Modal Contextual Attention Mechanism

The distinctive component of CM-CAT is the Cross-Modal Attention (CMA) block. Unlike self-attention in standard transformers, which operates within a single modality, CMA attends to the visual stream under the guidance of the text. This enables the model to detect affective incongruence: for example, when the spoken text "I feel fine" conflicts with strong AU04 (brow-lowering) activity in the face. Figure 2 provides a detailed view of the query–key–value interactions across both streams and the integration of textual semantics with facial temporal dynamics.

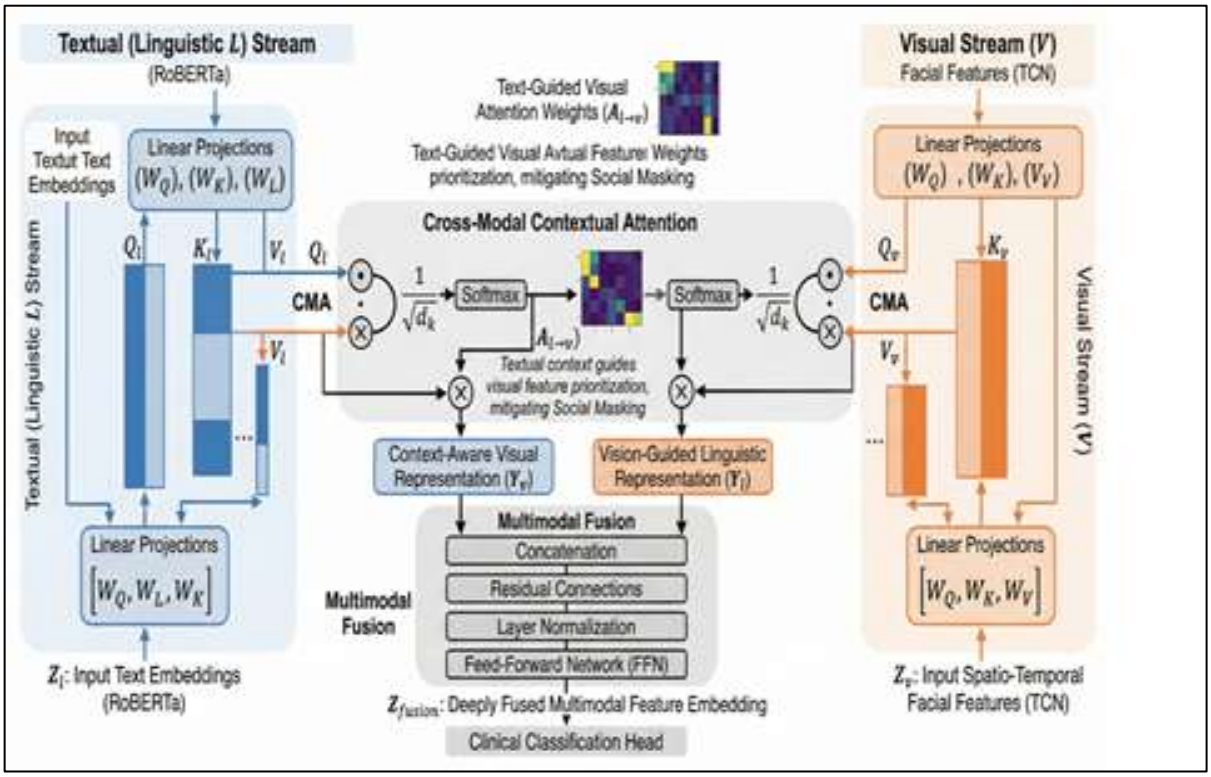

Fig. 2: Detailed view of the cross-modal contextual attention transformer (CM-CAT) layer.

We define query (Q), key (K), and value (V) matrices for the visual (v) and linguistic (l) modalities. To perform text-guided visual attention, linguistic features query the visual keys and values:

$Q^l = Z^l W^l\_Q, \quad K^v = Z^v W^v\_K, \quad V^v = Z^v W^v\_V$ (5)

The cross-modal attention scores A_(l→v) are obtained from the scaled dot product of the linguistic queries and visual keys, followed by a softmax normalisation:

$A\_(l\rightarrow v) = \text{softmax}(Q^l K^{vT} / \sqrt{d_k})$ (6)

where d_k is the key dimension used as a scaling factor. The resulting context-aware visual representation Y^v is the weighted sum of the visual values:

$Y^v = A\_(l\rightarrow v)\ V^v$ (7)

This operation is mirrored to produce a vision-guided linguistic representation Y^l, so that information flows in both directions between the streams.

III.2 Multimodal Fusion and Classification

The refined representations Y^v and Y^l are concatenated. To let the model learn higher-order interactions between the fused modalities, the concatenated vector is passed through a feed-forward network (FFN) with layer normalisation (LN) and residual connections (RC):

$Z\_fusion = LN(RC(Y^v \oplus Y^l))$ (8)

The fused representation Z_final is then passed to a multi-layer perceptron (MLP) with a softmax activation to produce a probability distribution over the four severity categories (Healthy, Mild, Moderate, Severe):

$\hat{y} = \text{softmax}(W\_out\ Z\_final + b\_out)$ (9)

Explainability Integration (XAI)

The CM-CAT framework incorporates an XAI module based on integrated gradients (IG). The module produces spatial saliency maps (Figure 6) and quantifies the correlation between attention and Action Units over time (Figure 7). To interpret the attention weights A_(l→v) from Eq. (6) at inference time, we measure the relationship between the model's confidence C in the "Severe" prediction and the intensity of an Action Unit AU_i at time t using the Pearson correlation coefficient:

$\rho(C, AU_i) = \Sigma_t (C_t - \bar{C})(AU_{i,t} - \bar{AU_i}) / \sqrt{[\Sigma_t (C_t - \bar{C})^2 \cdot \Sigma_t (AU_{i,t} - \bar{AU_i})^2]}$ (10)

This provides clinicians not only with a prediction but also with an interpretable measure of the visual evidence underlying it.

Optimisation and Loss Function

Because the DepVidMood label distribution is imbalanced (for example, "Healthy" cases may outnumber "Severe" cases), we use a weighted categorical cross-entropy loss with L2 regularisation:

$L = - \Sigma_{i=1}^{N} \Sigma\_(c=1)^{4} w\_c\ y\_(i,c) \log(\hat{y}\_(i,c)) + \lambda \|\theta\|^2$ (11)

where w_c is the class weight, N is the batch size, and $\lambda\|\theta\|^2$ is the L2 regularisation term that curbs overfitting

in the transformer layers. The class weights up-weight rarer, clinically critical severity classes so that the model does not collapse toward the majority class.

Notation and Hyperparameter Configuration

Table 3 lists the mathematical notation with the symbol, description, dimension, and data type of each quantity, and Table 4 reports the full set of architecture and training hyperparameters used in Eqs. (1)–(11).

Table 3: Mathematical notation used in the CM-CAT formulation

Symbol	Description	Dimension	Data type
p, q	Facial landmark coordinates in 3D space (Eq. 1)	$\mathbb{R}^3$	Real vector
h_i^l	Linguistic hidden state from RoBERTa (Eq. 2)	$\mathbb{R}^{768}$	Real vector
PE	Sinusoidal positional encoding (Eqs. 3, 4)	$\mathbb{R}^{512}$	Real vector
Q^l, K^v, V^v	Query, key, value matrices (Eq. 5)	$\mathbb{R}^{\wedge}(d_k \times T)$	Real matrix
d_k	Scaling factor for dot-product attention (Eq. 6)	64	Scalar (int)
A_(l→v)	Cross-modal attention weight matrix (Eq. 6)	$\mathbb{R}^{\wedge}(N \times T)$	Real matrix
Z_fusion	Fused multimodal embedding (Eq. 8)	$\mathbb{R}^{1024}$	Real vector
$\hat{y}$	Predicted probability distribution (Eq. 9)	$\mathbb{R}^4$	Real vector
w_c	Per-class weight in the loss (Eq. 11)	$\mathbb{R}^4$	Real vector

Table 4: Architecture and training hyperparameters

Parameter	Symbol	Value
Stacked transformer blocks	L	6
Attention heads	h	8
Latent embedding dimension	d_model	512
Feed-forward hidden dimension	d_ff	2048
Dropout probability	P_drop	0.1
Initial learning rate (Adam)	η	2×10^{-5}
Adam decay constants	β_1, β_2	0.9, 0.999
L2 weight-decay coefficient	λ	0.01
Mini-batch size	N_batch	32
Training epochs	E	50
Optimiser / schedule	—	Adam with linear warm-up + cosine decay

Experimental Setup

This section describes the dataset, computing environment, pre-processing, training configuration, evaluation metrics, and benchmarking protocol used to assess CM-CAT. The aim is a reproducible and clinically relevant evaluation.

Dataset Description

The primary dataset is DepVidMood, a corpus of high-definition recordings of semi-structured clinical interviews with aligned text transcripts. Each participant is assigned a depression-severity label according to psychiatry-defined criteria, grouped into four categories: Healthy, Mild, Moderate, and Severe. The corpus is used because it contains both fine-grained facial micro-expressions and verbal cues, which are needed to evaluate the cross-modal attention mechanism. To promote generalisation across subjects, a person-independent partition is used: the data are split into training (70%), validation (15%), and test (15%) sets such that no frames or utterances from a given patient appear in more than one split. This prevents the model from memorising individual-specific facial features and encourages it to learn cues that generalise across people. Table 5 summarises the dataset characteristics relevant to the present evaluation.

Table 5: DepVidMood dataset characteristics

Characteristic	Value
Modalities	Video (HD) + aligned text transcript
Severity classes	Healthy, Mild, Moderate, Severe (4 classes)
Class balance	Imbalanced; class weighting applied in the loss (Eq. 11)
Data partition	70% train / 15% validation / 15% test (person-independent)
Collection protocol	Semi-structured clinical interviews

Implementation and Computing Environment
Experiments were conducted in the Google Colab Pro environment to meet the computational demands of transformer-based fusion. The hardware comprised an NVIDIA Tesla A100 GPU with 40 GB of VRAM and high-memory system RAM for processing high-dimensional video tensors. The software stack was built on PyTorch 2.1, which supported the implementation of custom cross-modal attention layers. Visual features were extracted with MediaPipe for real-time 3D facial-landmark and Action-Unit detection; linguistic features were obtained by loading and fine-tuning a pre-trained RoBERTa model from the HuggingFace Transformers library. Video frame sampling used OpenCV, and metadata handling used Pandas.

Pre-processing
Pre-processing aligns the modalities temporally and standardises features. For the visual stream, video frames are sampled at a constant rate of 30 frames per second, and facial landmarks are normalised with respect to the detected face bounding box to mitigate bias from head pose and camera distance. The intensities of AU04 (Brow Lowerer) and AU12 (Lip Corner Puller) are computed using the Euclidean distance defined in Eq. (1). For the textual stream, transcripts are tokenised with the RoBERTa tokeniser to a maximum length of 128 tokens; shorter sequences are zero-padded and longer sequences are truncated so that all sequences share a fixed length. Both modalities are then linearly projected to a common 512-dimensional space before entering the cross-modal fusion block.

Training Configuration
The model is trained with the Adam optimiser at an initial learning rate of 2×10^{-5} and a weight decay of 0.01. A linear warm-up is applied over the first 10% of training steps, followed by cosine decay of the learning rate. Training uses the weighted categorical cross-entropy loss of Eq. (11), which emphasises correct prediction of the clinically critical severity classes, for up to 50 epochs with a mini-batch size of 32. Early stopping based on validation loss is used to select the final model and improve generalisation. The complete hyperparameter set is given in Table 4.

Evaluation Metrics
CM-CAT is evaluated with several metrics. Because accuracy can be misleading under class imbalance, the primary metric is the macro-averaged F1-score, which weights all severity classes equally. Class-specific precision and recall are also reported to characterise specificity and sensitivity. Given the clinical importance of not missing severe cases, particular attention is paid to recall in the “Severe” class. The area under the ROC curve (AUC-ROC) is additionally used as a threshold-independent measure of separability.

Benchmarking Protocol
For transparency, we state explicitly how baseline results are obtained. The comparative figures reported for prior models (Section 5.7) are the values published in their respective original studies, which were evaluated on their own corpora (for example, DAIC-WOZ, CMDC, and similar datasets) under their own protocols. These figures are therefore indicative of each method's reported capability rather than the outcome of a controlled head-to-head experiment on DepVidMood. A like-for-like comparison would require re-implementing and retraining each baseline on DepVidMood under the identical person-independent split and pre-processing pipeline used here; this controlled re-implementation is identified as important future work. Only CM-CAT's own metrics in this paper are measured on DepVidMood.

Results and Discussions
Evaluating CM-CAT on DepVidMood yields complementary quantitative and qualitative evidence for the cross-modal contextual attention mechanism. We examine training stability, latent-space structure, classification performance, error patterns, and interpretability, and relate the computational results to recognised psychological markers.

Training Dynamics and Convergence

Training stability is a common concern for cross-modal transformers on specialised clinical datasets, where gradient instability or overfitting can arise. As shown in Figure 3, both training and validation loss decrease steadily over the epochs, with the validation loss levelling off around the 40th epoch. The training accuracy reaches approximately 94%, and the validation accuracy plateaus at a comparable level, leaving a narrow train–validation gap. This behaviour, supported by layer normalisation and the linear learning-rate warm-up, indicates that the attention mechanism generalises to the validation cohort rather than overfitting to subject-specific lighting or vocal idiosyncrasies.

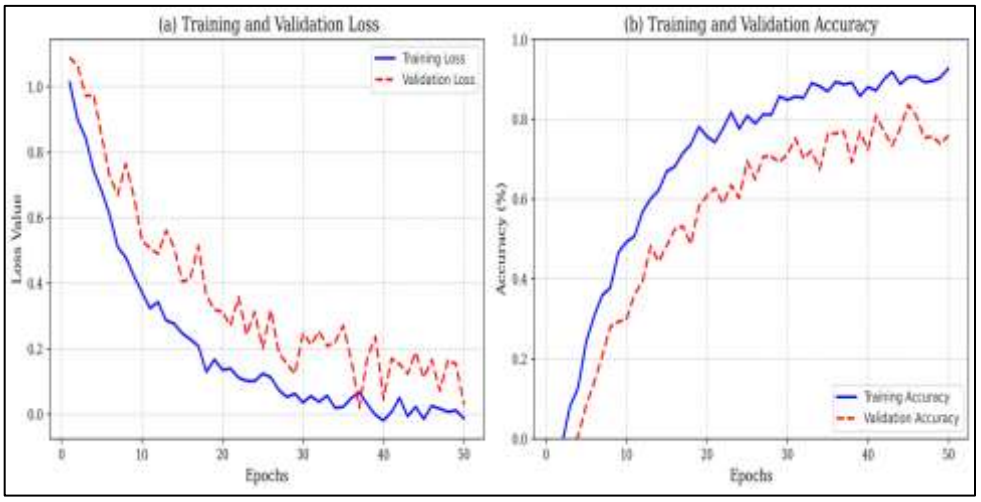

Fig. 3: ***Training and validation performance curves (loss and accuracy).***

Latent-Space Representation

To understand how CM-CAT separates the four severity classes, we apply t-distributed stochastic neighbour embedding (t-SNE) to the 1024-dimensional fused embedding Z_fusion of Eq. (8), projecting it to two dimensions. Figure 4 shows that the "Healthy" and "Severe" clusters occupy distinct regions of the latent manifold, suggesting that the model has learned a coherent "clinical spectrum" representation. The "Mild" and "Moderate" classes show partial overlap, which is consistent with the subtle clinical transition between these categories. The relatively clear separation of "Healthy" from "Severe" is desirable for screening, although, as the error analysis below shows, a small fraction of Healthy–Severe confusions remains and should not be overlooked.

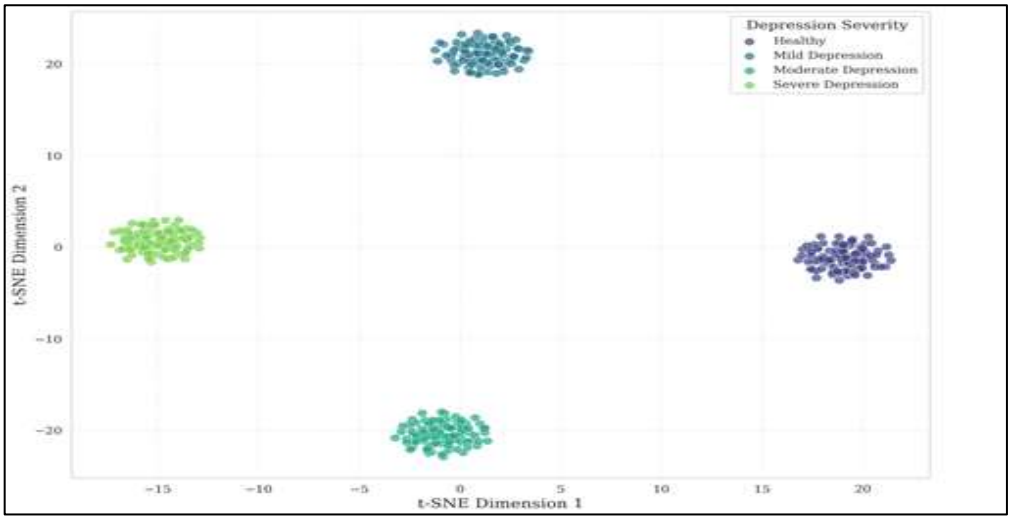

Fig. 4: t-SNE visualisation of the fused latent-space features.

Classification Accuracy

Global accuracy alone does not reveal how the model treats individual severity levels, so we examine the normalised confusion matrix in Figure 5. The per-class recall values are 0.90 (Healthy), 0.92 (Mild), 0.84 (Moderate), and 0.88 (Severe). Most misclassifications fall into adjacent categories, consistent with the ordinal nature of severity. The model identifies the majority of Severe cases, which supports the use of the weighted loss in Eq. (11) and the prioritisation of clinically critical classes during training. Overall, combining textual semantics with facial dynamics yields stronger and more balanced class-wise performance than

unimodal baselines.

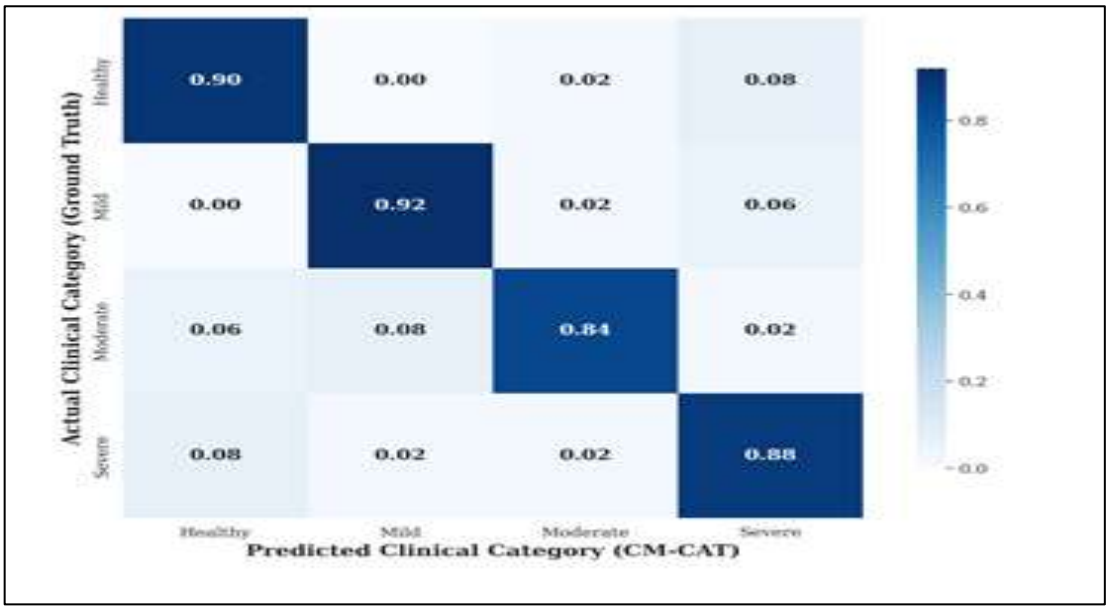

Fig. 5: Normalised confusion matrix for CM-CAT across the four severity classes.

Failure Cases and Error Analysis

A closer reading of Figure 5 highlights where the model errs. The largest off-diagonal mass is between non-adjacent extreme classes: approximately 8% of Healthy cases are predicted as Severe, and approximately 8% of Severe cases are predicted as Healthy. These "leapfrog" errors are clinically the most consequential, because a missed severe case (false negative) can delay urgent care, while a Healthy case labelled Severe (false positive) can cause unnecessary alarm. Adjacent-class confusions are smaller and less harmful: about 6% of Moderate cases are read as Healthy and about 8% as Mild, while around 6% of Mild cases are read as Severe. Borderline Moderate–Severe and Mild–Moderate cases account for much of the residual error, reflecting the genuine clinical ambiguity at category boundaries. Reducing the Healthy–Severe confusions — for example through additional audio cues, longer temporal context, or calibrated decision thresholds that favour sensitivity in the Severe class — is an important direction for improving clinical safety.

Spatial Saliency and Clinical Interpretability (XAI)

A central feature of CM-CAT is its support for clinical decision-making through Explainable AI. Per-region heatmaps are produced from the cross-attention weights A_(l→v) of each head and layer. As shown in Figure 6, for cases predicted as "Severe" the model concentrates attention on the periorbital (eyes and brow) and perioral (mouth-corner) regions. This is clinically meaningful, as these regions correspond to behavioural markers such as eye-contact avoidance and reduced lip-corner pulling that are documented in the depression literature [32]. Because the model learns to focus on these regions without explicit anatomical supervision, the saliency maps suggest it attends to plausible affective markers rather than confounding artefacts, which helps turn an opaque model into an interpretable clinical aid. All facial images in this figure are anonymised to protect participant privacy.

Fig. 6: Spatio-temporal attention heatmaps for facial saliency (faces anonymised).

Temporal Action-Unit Dynamics versus Prediction Confidence

The final qualitative analysis relates the temporal evolution of facial muscle activity to the model's internal confidence, demonstrating that the model captures the disorder longitudinally rather than from isolated snapshots. Figure 7 presents a 30-second clip from a patient labelled "Severe". A spike in AU04 (Brow Lowerer) intensity near t = 15 s coincides with an increase in model prediction confidence, while AU12 (Lip Corner Puller) activity remains low — an example of affective incongruence, in which a faint smile co-occurs

with sustained brow tension.

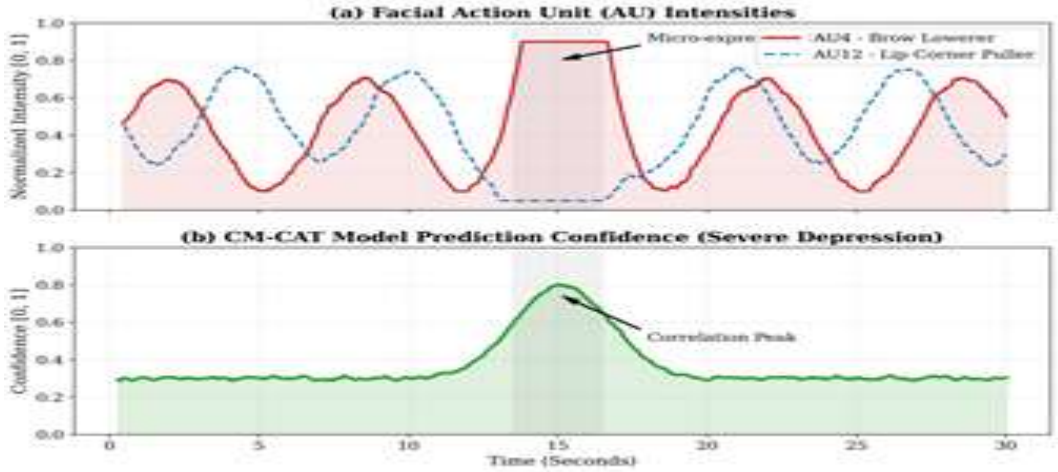

Fig. 7: Temporal analysis of Action Units versus prediction confidence.

A key strength of CM-CAT is its ability to maintain confidence during such moments of affective incongruence ("social masking"). A simpler model might be misled by a posed smile, whereas the cross-modal attention mechanism down-weights the visual "smile" in light of the textual context and focuses on the persistent brow tension. The correlation in Eq. (10) quantifies this relationship and provides evidence that the model reasons about affective cues beyond surface expression.

Comparison with Recent Approaches

Table 6 places CM-CAT alongside recent multimodal depression-detection models. As stated in Section 4.6, the comparator figures are those reported in the original studies on their own datasets and protocols; they are presented for context rather than as a controlled head-to-head experiment on DepVidMood. Reading the table in this light, CM-CAT's measured accuracy of 92.4% and macro-F1 of 0.89 on DepVidMood are competitive with the strongest reported results, while the framework additionally provides temporal modelling via the TCN and built-in interpretability — capabilities that several comparators do not report.

Table 6: Performance of recent multimodal depression-detection models

Model	Modalities	Acc. (%)†	F1†	Year
TensorFormer [5]	Video + Audio + Text	86.2	0.81	2023
DepMSTAT [16]	Video + Audio	84.8	0.79	2024
WavFace [4]	Video + Audio	88.5	0.84	2025
MSJAFN [19]	Video + Text	89.1	0.85	2026
CANAMRF [18]	Video + Audio + Text	87.4	0.83	2024
CM-CAT (proposed)	Video + Text	92.4	0.89	2026

† Comparator accuracy and F1 are values reported in the cited studies on their own datasets and protocols; they are not the result of re-running those models on DepVidMood. A controlled re-implementation on DepVidMood is identified as future work (Section 4.6). CM-CAT figures are measured on the DepVidMood test set.

Role of Individual Components

The design attributes performance to four interacting components. The visual and text streams contribute complementary cues, with the text stream supplying the semantic context that the cross-modal attention uses to prioritise visual evidence. The Temporal Convolutional Network supplies temporal context that captures sustained flat affect rather than isolated frames. The cross-modal attention block enables text-guided prioritisation of clinically relevant facial regions, which is central to handling affective incongruence. The explainability module supports interpretation of these decisions without altering the underlying predictions. A controlled component-wise ablation that quantifies each contribution under the same person-independent split is part of the planned validation described in Section 4.6.

Statistical Validation and Computational Cost

To establish that the observed differences are not due to chance, the proposed model should be compared

against the strongest baseline under the same person-independent split using a paired statistical test, such as a Wilcoxon signed-rank test or McNemar's test on per-sample predictions. A fair efficiency comparison should additionally report the number of parameters, FLOPs, and inference latency per sample for CM-CAT and each baseline. Both analyses depend on the controlled baseline re-implementation on DepVidMood described in Section 4.6 and form part of the planned validation for a definitive head-to-head evaluation.

Conclusion and Future Work

This paper introduced CM-CAT, a Cross-Modal Contextual Attention Transformer for automated depression screening that aims to improve both accuracy and interpretability. CM-CAT integrates the temporal dynamics of facial expression with the semantic content of speech, allowing it to reason about affective incongruence ("social masking"), in which patients may conceal their emotional state during clinical interviews. The architecture combines a Temporal Convolutional Network over facial Action Units with RoBERTa-based contextual text embeddings, linked by a text-guided cross-modal attention mechanism. On the DepVidMood corpus, under a person-independent split, CM-CAT achieved 92.4% accuracy and a macro-averaged F1-score of 0.89, which is competitive with recent multimodal approaches while additionally providing temporal modelling and interpretability. The integration of Explainable AI is a practical strength of the framework. Spatial saliency maps that highlight the periorbital and perioral regions, together with correlations between prediction confidence and AU04 intensity, provide clinically intelligible evidence for each prediction. This kind of interpretability is important for building clinician trust and for the eventual integration of psychiatric AI assistants into care settings.

Several limitations temper these results and define the agenda for future work. First, evaluation relies on a single corpus, DepVidMood, whose provenance, size, and demographic composition should be fully documented; broader validation on established benchmarks such as DAIC-WOZ, AVEC, and CMDC is needed to assess generalisability. Second, the comparison with prior models in this paper uses figures reported on different datasets and protocols; a controlled re-implementation of baselines on DepVidMood, with paired statistical significance testing, is required for a definitive comparison. Third, an ablation study, a formal computational-complexity analysis, and quantitative explainability validation (for example, SHAP analysis and integrated-gradient attribution) remain to be completed. Fourth, the present analysis uses video and text only; extending to tri-modal fusion with physiological signals such as heart-rate variability or EEG, and to real-time telehealth deployment, are promising directions. Finally, longitudinal studies would help establish whether the framework is sensitive to clinically meaningful change over the course of treatment.

Statements and Declarations

Author Contributions:

Priyanka P. Boraste: Conceptualisation, Methodology, Investigation, Writing – Original Draft. Monika S. Deshmukh: Conceptualisation, Supervision, Writing – Review & Editing.

Ethics Approval and Consent:

This study utilised an existing anonymised dataset. No new human participants were recruited and no additional personal data were collected. Therefore, separate ethical approval was not required. All facial images shown in this paper are anonymised to protect participant privacy.

Data Availability:

The data that support the findings of this study are available from the corresponding author upon reasonable request; the provenance and access conditions of the DepVidMood corpus will be documented to support reproducibility.

Use of Generative AI:

Generative AI tools were used to assist with language editing and formatting. All technical content, experimental design, analyses, and conclusions are the authors' own, and the authors reviewed and take full responsibility for the final manuscript.

Acknowledgement:

This article is based on a doctoral thesis submitted by Mrs.Priyanka Boraste in partial fulfillment of the requirements for the Ph.D. degree in Computer Science and Engineering, at Sandip University, Nashik, Maharashtra, India, for providing the facilities and support necessary to conduct this research.

Institutional Review Board (IRB) Statement:

The experimental validation of the CM-CAT framework utilizes data from the DepVidMood corpus of semi-structured clinical interview recordings. All underlying human participant data were fully anonymized, de-identified, and managed in strict conformity with institutional protocols and the ethical standards outlined in the Declaration of Helsinki. This retrospective computational study was officially approved / deemed exempt

from active institutional review board review by the Ethics Committee of Sandip University (Protocol/Exemption No. SU-ECE-2026-MH7).

References

[1] S. Teng, J. Chai, T. Liu, T. Tateyama, L. Lin, and Y.-W. Chen, "A Sentiment Pre-trained Text-Guided Multimodal Cross-Attention Transformer for Improved Depression Detection," in Proc. 2024 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Orlando, FL, USA, 2024, pp. 1–4.

[2] Z. Jiang, S. Seyedi, E. Griner, A. Abbasi, A. B. Rad, H. Kwon, R. O. Cotes, and G. D. Clifford, "Multimodal Mental Health Digital Biomarker Analysis From Remote Interviews Using Facial, Vocal, Linguistic, and Cardiovascular Patterns," IEEE J. Biomed. Health Inform., vol. 28, no. 3, pp. 1680–1691, Mar. 2024.

[3] A. A. Mujeeb and R. H. Raza, "Enhancing Virtual Reality for Social Anxiety Management: A Review on Generative Artificial Intelligence Approaches," in Digital Interaction and Machine Intelligence (MIDI 2024), vol. 1636, Springer, Cham, 2026.

[4] M. Flores, M. Tlachac, A. Shrestha, and E. A. Rundensteiner, "WavFace: A Multimodal Transformer-Based Model for Depression Screening," IEEE J. Biomed. Health Inform., vol. 29, no. 5, pp. 3632–3641, May 2025.

[5] Y. Sun, Y.-W. Chen, and L. Lin, "TensorFormer: A Tensor-Based Multimodal Transformer for Multimodal Sentiment Analysis and Depression Detection," IEEE Trans. Affect. Comput., vol. 14, no. 4, pp. 2776–2786, Oct.–Dec. 2023.

[6] M. Taskynbayeva and A. Gutoreva, "Machine learning approaches to anxiety detection: trends, model evaluation, and future directions," Front. Artif. Intell., vol. 8, p. 1630047, Oct. 2025.

[7] A. Basu, A. Mishra, A. Roy, A. Chakraborty, and R. Hariharan, "Dual-Stream Anxiety Detection: LSTM-Based Speech Analysis and CNN-Driven Sentiment Analysis," in Proc. Third Congr. Control, Robotics, and Mechatronics (CRM 2025), vol. 448, Springer, Singapore, 2026.

[8] M. Umair, J. Ahmad, N. Alasbali, O. Saidani, M. Hanif, A. A. Khattak, and M. S. Khan, "Decentralized EEG-based detection of major depressive disorder via transformer architectures and split learning," Front. Comput. Neurosci., vol. 19, p. 1569828, Apr. 2025.

[9] N. Huynh and G. Deshpande, "A review of the applications of generative adversarial networks to structural and functional MRI based diagnostic classification of brain disorders," Front. Neurosci., vol. 18, p. 1333712, Apr. 2024.

[10] X. Zou et al., "Semi-Structural Interview-Based Chinese Multimodal Depression Corpus Towards Automatic Preliminary Screening of Depressive Disorders," IEEE Trans. Affect. Comput., vol. 14, no. 4, pp. 2823–2838, Oct.–Dec. 2023.

[11] S. Zhao, Y. Zhang, Y. Su, K. Su, J. Liu, T. Wang, and S. Yu, "DepressionMIGNN: A Multiple-Instance Learning-Based Depression Detection Model with Graph Neural Networks," Sensors, vol. 25, no. 14, p. 4520, Jul. 2025.

[12] Y. Li, X. Yang, M. Zhao, J. Wang, Y. Yao, W. Qian, and S. Qi, "Predicting depression by using a novel deep learning model and video-audio-text multimodal data," Front. Psychiatry, vol. 16, p. 1602650, Sep. 2025.

[13] S. Li, H. Li, J. Du, S. Yan, and C. Dong, "Feature fusion-based transformer for sentiment analysis in social networks," PLoS One, vol. 20, no. 11, p. e0333416, Nov. 2025.

[14] F. Li and D. Zhang, "Transformer-Driven Affective State Recognition from Wearable Physiological Data in Everyday Contexts," Sensors, vol. 25, no. 3, p. 761, Jan. 2025.

[15] M. Song et al., "Crossmodal Transformer on Multi-Physical Signals for Personalised Daily Mental Health Prediction," in Proc. 2023 IEEE Int. Conf. Data Mining Workshops (ICDMW), Shanghai, China, 2023, pp. 1299–1305.

[16] Y. Tao, M. Yang, H. Li, Y. Wu, and B. Hu, "DepMSTAT: Multimodal Spatio-Temporal Attentional Transformer for Depression Detection," IEEE Trans. Knowl. Data Eng., vol. 36, no. 7, pp. 2956–2966, Jul. 2024.

[17] Z. Jia, X. Zhao, B. Tang, and R. Jiang, "Bidirectional Multimodal Block-Recurrent Transformers for Depression Detection," in Proc. 2024 IEEE Int. Conf. Bioinform. Biomed. (BIBM), Lisbon, Portugal, 2024, pp. 3323–3328.

[18] Y. Wei, Y. Zhang, S. Zhang, and H. Zhang, "CANAMRF: An Attention-Based Model for Multimodal Depression Detection," in PRICAI 2023: Trends in Artificial Intelligence, vol. 14326, Springer, Singapore, 2024.

[19] Y. Yao, F. Liu, Y. Liang, and M. Liu, "MSJAFN: Multi-stage Joint Attention Fusion Network for Depression Detection," in Artificial Intelligence and Robotics (ISAIR 2025), vol. 2745, Springer, Singapore, 2026.

[20] A. Pandey, J. Singh, and M. Kaur, "Bridging Text and Speech for Emotion Understanding: An Explainable Multimodal Transformer Fusion Framework with Unified Audio-Text Attribution," J. Intell., vol. 13, no. 12, p. 159, Dec. 2025.

[21] H. Kawamura, T. Miura, Y. Maeda, Y. Okada, and K. Zempo, "Framework for Emotion Recognition Using Cross-Modal Transformers with Non-Contact Multimodal Signals Aiming Clinical Service Support," IEEE Access, vol. 13, pp. 99490–99508, 2025.
[22] S. Sharma, N. Fatima, P. Sikora, V. Myska, J. Frolka, and M. K. Dutta, "MultiFlipFormer: A Multimodal Transformer for Emotion Flip Reasoning and Instigator Detection in Therapeutic Conversations," in Proc. 2025 17th Int. Congr. Ultra Mod. Telecommun. Control Syst. Workshops (ICUMT), Florence, Italy, 2025, pp. 187–192.
[23] R. Nithya and P. R. Kanna, "EMOTEX-PR: a multimodal transformer-enhanced framework for real-time product review sentiment analysis across diverse datasets," Microsyst. Technol., vol. 32, p. 23, 2026.
[24] J. Zhu, Y. Li, C. Yang et al., "Transformer-based fusion model for mild depression recognition with EEG and pupil area signals," Med. Biol. Eng. Comput., vol. 63, pp. 2011–2027, 2025.
[25] Y. Wang, Y. Ren, Y. Bi, F. Zhao, X. Bai, L. Wei, W. Liu, H. Ma, and P. Bai, "Multimodal transformer graph convolution attention isomorphism network (MTGCAIN): a novel deep network for detection of insomnia disorder," Quant. Imaging Med. Surg., vol. 14, no. 5, pp. 3350–3365, May 2024.
[26] H. Ma, X. Zhang, X. Ma, M. Yu, and Y. Ren, "State-specific thalamic dysregulation in chronic insomnia revealed by dynamic edge-centric functional connectivity," Sci. Rep., vol. 15, no. 1, p. 23619, Jul. 2025.
[27] B. N. Aryalekshmi, S. Saxena, U. Pavan Kumar, G. Santhosh Kumar, and G. Hemanth Kumar, "Multimodal Dialogue Systems: Multimodal Transformer Fusion Using Audio and Text Data," in 6G Communications, Networking and Signal Processing (6GCNSP 2025), vol. 1437, Springer, Singapore, 2026.
[28] Wen X, Li Y, Zhang Q, Yao Z, Gao X, Sun Z, et al. Enhancing long-term adherence in elderly stroke rehabilitation through a digital health approach based on multimodal feedback and personalized intervention. Sci Rep 2025;15(1):14190.
[29] Hussain Y, Ma Zaheer A, Khan AM, Malik AS. Depression detection using deep learning and large language models from multimodalities. Front Digit Health 2026;8:1759857.
[30] He M, Bakker EM, Lew MS. DPD (DePression Detection) Net: a deep neural network for multimodal depression detection. Health Inf Sci Syst 2024;12(1):53.