Journal of Artificial Intelligence and Technology, 2026, 6, 665-676
https://doi.org/10.37965/jait.2026.1086

ISTP
RESEARCH ARTICLE

DiscRST-HTT: Hierarchical Multimodal Transformer with Adaptive Gating for Discourse-Aware Depression Detection

K. S. Srinath,[1] K. Kiran,[1] B. Naveen Kumar,[2] P. Deepa Shenoy,[1] and K. R. Venugopal[3]
[1]Department of Computer Science and Engineering, University of Visvesvaraya College of Engineering, Bengaluru, India
[2]Department of Computer Science and Engineering (Data Science), Atria Institute of Technology, Bengaluru, India
[3]Bangalore University, Bengaluru, India

(Received 15 December 2025; Revised 17 May 2026; Accepted 26 May 2026; Published online 01 July 2026)

Abstract: Recently, depression detection through social media activity has emerged as an effective and gain insights to understand the mental health condition of a user. However, traditional transformer-based models primarily focus on textual semantics and neglect the temporal evolution of behavior, which limits their adaptability and reliability. To address these challenges, a Discourse-aware Rhetorical Structure Theory with Hierarchical Temporal Transformer (DiscRST-HTT) is proposed for integrating contextual, rhetorical, and behavioral cues for comprehensive depression analysis. The contextual semantics are captured through a Mental Bidirectional Encoder Representations from Transformer (MentalBERT), while rhetorical embeddings are derived from a Neural Rhetorical Structure Theory (NeuralRST) model. In contrast, the model learns behavioral embeddings, which are captured based on posting frequency. These embeddings are fused through a multimodal adaptive gating mechanism (MAGM) by dynamically balancing modality contributions. Then, the fused embeddings are processed by a hierarchical temporal transformer (HTT) model, which captures both intra-post and inter-post dependencies with an ordinal regression head to learn ordered depression severity representations. While the final evaluation is conducted through binary depression classification derived from threshold ordinal output. From the results, the proposed DiscRST-HTT attains better binary classification performance derived from ordinal severity predictions achieving accuracy of 98.75%, precision 98.34%, recall 99.05%, and F1-score 98.78% on the Reddit dataset when compared to the existing BERT with convolutional neural network (BERT-CNN) model.

Keywords: depression detection; hierarchical temporal transformer; multimodal adaptive gating mechanism; rhetorical structure theory; social media

I. INTRODUCTION

Over the years, depression has become one of the most dominant mental health disorders globally, and timely identification remains crucial for early intervention [1,2]. The rapid expansion of online communities such as Reddit and Twitter users often disclose emotions and experiences, which reflect their mental well-being [3]. This makes social media a rich and unremarkable resource for automated depression detection with natural language processing (NLP) [4]. Recent advancements in deep learning (DL) and transformer architectures enabled models to learn semantic, emotional, and contextual cues from text data, which provides a promising accuracy when compared to traditional lexical and psychological approaches [5]. Nevertheless, linguistic diversity and contextual variability limit reliable prediction [6]. Despite advancements, the existing social media depression detection system faces several unaddressed challenges [7]. Most transformer-based models consider each post independently and do not consider the user's behavioral and emotional change over time [8]. In addition, the current model primarily concentrates on surface-level semantics, overlooking rhetorical structure such as how ideas and emotions are logically connected within text data [9]. Moreover, a major concern is the lack of interpretability and calibrated confidence, which undermines the reliability of predictions in sensitive clinical settings [10]. These limitations highlight the integration of linguistic content with discourse awareness and behavioral activity for a temporal-aware and explainable framework [11].

These limitations motivate the observation of depression detection, which is not only through word choice or sentiment but also through how thoughts are structured and when they are expressed [12]. Depressed users sometimes share their thoughts on social media in discourse patterns such as frequent contrastive reasoning and post at irregular times [13]. Therefore, capturing rhetorical coherence and behavioral regularities alongside textual semantics provides a deeper psychological insight [14]. Although earlier studies have merged linguistic, sentiment, or behavioral features using a simple fusion mechanism, they ignore the deeper rhetorical structure and temporal progression of user expression over time. The existing transformer-based or calibrated models operate at the post-level without modeling the depressive language that evolves over time with discourse relations [15]. The state-of-the-art methods have considered recent advancements including deep multimodal and sentiment-guided architectures for depression detection. A sentiment-guided transformer with severity-aware contrastive learning (STCL) is introduced to effectively distinguish depression severity through contextual and emotional

Corresponding author: K. S. Srinath (e-mail: solansrinath@gmail.com).

representations [16]. In addition, a cross-attention multimodal fusion (CAMF) model is introduced to integrate textual and behavioral cues with attention-based feature interaction for improving precision [17]. Moreover, combined linguistic and behavioral insights are utilized to capture psychological tendencies from social media data [18]. All these models achieve strong performance but lack integration of rhetorical structure and temporal discourse modeling, which are the key components to the proposed Discourse-aware Rhetorical Structure Theory with Hierarchical Temporal Transformer (DiscRST-HTT) framework. Moreover, the proposed model uses Mental Bidirectional Encoder Representations from Transformer (MentalBERT), Neural Rhetorical Structure Theory (NeuralRST), and behavioral cues independently, which are used in the prior works. These frameworks are typically employed in a simpler feature concatenation without modeling their interdependence. But the proposed model uses these models with hierarchical integration using discourse-aware, semantic, and behavioral representations. Here, the primary work of the proposed model is deploying an attention mechanism, which dynamically regulates modality contributions instead of assigning uniform weights. The transformer model effectively captures intra-post and inter-post emotional evaluation of a user. This unified design provides a deeper linguistic-behavioral coupling while improving temporal reliability. Even though the prior works incorporated semantic, sentiment, and behavioral cues, the models typically fuse these features at a coarse granularity, which often uses a simple concatenation or a static weighting without modeling the interaction between discourse structure and behavioral rhythms. Unlike these fusion models, the proposed model reflects how the rhetorical coherence, semantic tone, and posting irregularities evolve in depressive expression while providing a finer-grained multimodal representation. The proposed model performs ordinal severity-aware representation learning using Cumulative Ordinal Regression using a Logistic Link (CORAL)-based ordinal regression, while the final evaluation is conducted as binary depression classification for a fair comparison with the prior studies.

The contributions are listed as follows:

- The three isolated encoders collectively enhance the multimodal understanding by capturing the nuances with multiple models, including MentalBERT, NeuralRST, and activity patterns. These encoders provide a comprehensive linguistic-behavioral representation, which are essential for accurate depression detection.
- The multimodal adaptive gating mechanism (MGAM) dynamically learns the relative importance of contextual, rhetorical, and behavioral features while enabling selective information transmission across modalities. The adaptive fusion effectively reduces the redundancy and ensures balanced representation learning while improving depression detection and calibration.
- The hierarchical temporal transformer (HTT) model introduces a dual-level architecture that captures both intra-post discourse variations and inter-post temporal dependencies. The HTT model integrates time-aware attention and ordinal calibration to effectively model emotional progression in social media.

The remaining sections of this research paper are structured as follows: Section II describes the literature review, and Section III demonstrates the proposed methodology. Section IV presents the experimental results, and Section V presents the discussion. The conclusion of this research paper is provided in Section VI.

II. LITERATURE REVIEW

Ajmal *et al.* [19] developed a rhetorical structure theory and ensemble models (RSTFusionX) for depression prediction in social media posts. This RSTFusionX model analyzed the depressive language and attained rhetorical relations with rhetorical structure theory. Nevertheless, the RSTFusionX model was constrained with a high computational complexity and faced difficulty in accurately interpreting deep models, which affected reliability in depression prediction.

Chiong *et al.* [20] presented a text-based featuring approach for depression detection using machine learning (ML) classifiers and social media texts. The social media were preprocessed and extracted features to process with single and ensemble models for detecting depression without relying on explicit keywords. Nevertheless, the dependency on the dataset created a generalization issue, and it was limited to the nondepressive sample, which might affect detection accuracy across diverse data sources.

Qasim *et al.* [21] demonstrated the detection of depression severity in social media text using transformer-based models. The N-gram feature extraction was integrated with transformer embeddings and transformer-based models for depression severity classification. This model exhibited potential bias in social media text by limiting generalizability across various cultural or linguistic contexts.

Salameh *et al.* [22] represented a deep linguistic analysis for depression (DLAD) in social media using Robustly optimized Bidirectional Encoder Representations from Transformer algorithm (RoBERTa) and convolutional neural network (CNN). The DLAD was proposed to detect depression symptoms in social media posts. However, the model deployed to identify psychological cues showed potential bias in social media language across diverse user groups.

Ansari *et al.* [23] demonstrated an ensemble hybrid learning (EHL) method for automated depression detection. The EHL model achieved superior results by combining multiple features and classifiers. However, the model heavily depended on textual data quality and lacked consideration for multimodal signals such as speech or facial expressions.

Xin and Zakaria [24] illustrated an integrated BERT with CNN and bidirectional long short-term memory (BiLSTM) for explainable detection of depression in social media content. The model was fine-tuned with BERT-based models for depression detection with attention visualization methods. However, the dependency on data led to potential bias in social media content, which limited generalizability to diverse populations.

Ilias *et al.* [25] developed a calibration of transformer-based models for identifying stress and depression in social media. The calibration enhanced interpretability and reduced overfitting by representing it as a major advantage over conventional transformer models. However, the reliance on data availability and potential bias in social media posts showed generalizability issues across diverse populations and contexts.

A. RESEARCH GAP

Existing studies based on social media-based depression detection primarily focus on semantic or context text analysis using ML or transformer models but overlook the underlying rhetorical structure

and temporal evolution of user expressions. Even with high accuracy, the ensemble models and transformer-based methods often lack clear interpretability and well-calibrated confidence. Moreover, the recent advanced models analyze posts as isolated units while ignoring the progression of depressive cues across time. Therefore, there is a significant research gap in developing an integrated discourse-aware, temporally adaptive framework that can learn ordered depression severity representations while supporting the reliable binary depression detection.

B. PROBLEM STATEMENT

Despite notable progress in transformer-based depression detection, current approaches primarily focus on semantic or sentiment-level text understanding and overlook the rhetorical organization of thoughts with temporal evolution. Moreover, the behavioral activity patterns such as posting frequently or context utilization and more models lack interpretability and confidence calibration, which are essential for mental health applications. Consequently, the existing models fail to holistically represent users' linguistic and behavioral dimensions. Therefore, there is a need for a unified model that effectively captures temporal evolution and discourse-aware patterns while learning ordered depression severity representations that can be transformed into reliable binary depression predictions.

III. METHODOLOGY

In this section, the depression detection is employed with the proposed DiscRST-HTT model by integrating contextual, rhetorical, and behavioral information from social media. Initially, the user posts are preprocessed and grouped chronologically and then encoded by three specialized encoders. The encoders are used independently such as MentalBERT for extracting contextual semantics, NeuralRST for capturing rhetorical discourse relations, and behavioral encoder models for activity patterns such as posting frequency and engagement. These encoder representations are fused into a unified representation using a MAGM, which dynamically weights the modality importance. The fused features are then processed by an HTT, which learns both intra- and inter-pot dependencies to detect depression. The flow diagram of the proposed DiscRST-HTT model is provided in Fig. 1.

A. DATASET

The Reddit dataset is a social media collection that comprises posts from depressed and nondepressed users. The dataset contains 1841 users, where 1200 are positive users and 641 are negative users [26]. Reddit is a social media platform, which maintains the anonymity of its users, and it is widely used for discussing stigmatizing topics. The Reddit data is used to study the posts specifically from the Reddit users who write about mental health issues and who have proceeded to post topics about suicidal ideation. The data are randomly shuffled and split into train and test sets with an 80:20 split rate. The final frame of the dataset consists of a subreddit where the post is shared and the body consists of content posted by the user. In addition, the label specifies the category assigned to the post. If the post is normal, it is labeled as "0" and the depression post is indicated with a "0" label.

Consequently, the eRisk dataset is collected from the eRisk (Early Risk Prediction) forum, which is a public competition platform that facilitates multidisciplinary research and creates reusable datasets for assessing early risk detection in health and safety problem areas [27]. The eRisk 2018 dataset is developed to detect the early signs of depression. The dataset comprises 4498 depressed and nondepressed users, where 3728 users belong to the nondepressed and 770 belong to the depressed class. The data are concatenated and randomly shuffled and then split into 80:20 train and test split rate.

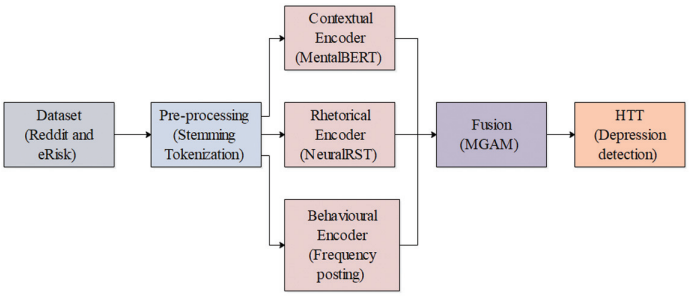

Fig. 1. The flow diagram of the proposed DiscRST-HTT for depression detection through social media.

B. DATA PREPROCESSING

The input data are first processed with NLP tools to preprocess before proceeding to the training. First, tokenization is employed to split the input posts into individual tokens. Then, punctuation and stop words are removed where the stemming is applied to reduce word length by making them into their root form. These preprocessing steps make the data clean and help the proposed model to group similar posts by user. The comments are converted into lowercase letters, and irrelevant text and extra white spaces are removed. Moreover, the comments are trimmed with regard to length, as short comments are particularly relevant to depression detection. Subsequently, all posts are grouped according to their corresponding users to preserve the temporal and contextual continuity of their activity. Each user is treated as an independent sequence to ensure that the chronological order of posts is maintained for subsequent temporal modeling. This grouping of posts by the user enables the model to analyze linguistic and behavioral evolution over time, thereby providing a coherent input structure for contextual, rhetorical, and temporal feature extraction. Consequently, to ensure the robustness for users with sparse activity or extremely short post histories, the proposed model adopts several stabilization mechanisms. The Z-score normalization analyzes the behavioral characteristics of users with low activity levels by scaling them relatively to the dataset mean and standard deviation. It reduces dominance of highly active users and allows low-activity user behaviors during model learning. This normalization ensures that variations in behavioral signals from low-activity users are preserved and effectively utilized by the model during training.

C. FEATURE EXTRACTION

The proposed model first extracts the contextual, temporal, and behavioral features from multiple encoders separately. First, the MentalBERT is employed to generate contextual embeddings.

1). CONTEXTUAL ENCODER. In the proposed model, the contextual encoder uses MentalBERT, which is a transformer designed and adapted for mental health text analysis. Unlike general-purpose models such as BERT and RoBERTa, MentalBERT is pretrained on large-scale Reddit and mental health forum datasets, enabling it

to capture the semantic meaning and psychological undertones of user posts. Each post P is first tokenized into subword units $\{t_1,t_2,\ldots,t_n\}$ using wordpiece tokenization. The attention mechanism for a given layer l specifies that query Q, key K, and value V matrices are generated from the input token embedding through a learned linear projection.

Here, $\mathrm{s} \in \mathbb{R}^d$ captures the meaning, context, and discourse flow of the post, specifies an attention score, and h_i denotes the contextual token representation. Moreover, an additional semantic projection head is employed to capture the emotional polarity in the post, which is applied on top of the semantic embedding e that is demonstrated in Equation (1):

$$e = \tanh(W_e \mathrm{s} + b_e) \tag{1}$$

Here, $W_e \in \mathbb{R}^{d\times d}$ and $b_e \in \mathbb{R}^d$ specify the trainable weight and bias parameters, respectively. The semantic projection head applies a nonlinear transformation on the pooled context embedding. For instance, $z \in \mathbb{R}^{model}$ refers to attention-weighted pooled embedding from the MentalBERT model. The activation function $\phi(\cdot)$ is defined as a Gaussian error linear unit (GELU), where the semantic projection is computed as $h = \phi(W_S z + b_s)$, $h \in\in \mathbb{R}^{d_s}$.

Furthermore, to derive a fixed-dimensional semantic embedding for the entire post, the MentalBERT model uses a pooling strategy where the hidden state is obtained through an attention-weighted pooling mechanism to highlight the important tokens, and it is depicted in Equation (2):

$$\mathrm{s} = \sum_{i=1}^{n} \alpha_i \mathrm{h}_i, \text{ where } \alpha_i = \frac{\exp(\mathrm{w}^\top \tanh(W\mathrm{h}_i))}{\sum_{j=1}^{n} \exp(\mathrm{w}^\top \tanh(W\mathrm{h}_j))} \tag{2}$$

2). RHETORICAL ENCODER. Several discourse parsing frameworks are available for RST including traditional parsers and neural-based approaches. However, in this work, the NeuralRST parser is employed due to its improved capability in capturing hierarchical discourse relations using neural representations. Compared with earlier rule-based or feature-engineered parsers, NeuralRST provides better generalization for informal social media text and produces more stable Elementary Discourse Unit (EDU) segmentation that is vital for reliable rhetorical feature extraction. Each EDU is encoded using the shared MentalBERT encoder to obtain semantic representations. Minimal EDUs with extremely low semantic content are mitigated through preprocessing strategies and attention weighting. During encoding, the attention mechanism assigns lower importance scores to discourse units with weak contextual embeddings to reduce the influence of noisy clause on the final representation. For each pair of EDUs (E_i,E_j), the rhetorical encoder predicts two attributes including a relation type r_{ij} and nuclearity n_{ij}, respectively. The parser constructs an RST tree with nodes and edges $T = (V,E)$ where node V denotes the EDUs and edges E represents the rhetorical relations r_{ij}. Here, each EDU is encoded using the same shared MentalBERT encoder to ensure that all discourse units lie in a consistent semantic space. No separate encoder is created for this, but instances are created; the parameter sharing is enforced across all EDUs post. Further, the encoder constructs a relation vector for each rhetorical relation connecting two EDUs with a relation type, which is represented in Equation (3):

$$r_{ij} = \tanh(W_r[h_i \| h_j] + b_r) \tag{3}$$

Where the term r_{ij} denotes the embedding of relation types for EDU pair (e_i,e_j). Each relation type is represented using a learned embedding vector, which is obtained from a trainable lookup table $E_{rel} \in \mathbb{R}^{|\mathcal{R}|\times d_{rel}}$.

The parser output is a categorical label, which is mapped to its corresponding embedding $r_{ij} = E_{rel}[r]$. This learned embedding allows the relation types to occupy continuous vector space instead of relying on fixed one-hot encodings while enabling the model to capture similarity patterns between rhetorical relations. The final relation vector r_{ij} is constructed from the concatenation of two EDU embeddings and the relation type embedding. Moreover, the concatenation highlights the importance of the nuclear unit by applying an attention weight that is provided in Equation (4):

$$\alpha_{ij} = \frac{\exp(u^\top r_{ij})}{\sum_{(p,q)\in E} \exp(u^\top r_{pq})} \tag{4}$$

where α_{ij} specifies an attention weight and $u^\top$ denotes a learnable context vector, r_{pq} specifies the relation type of pq, and the rhetorical embedding for the post is computed as the weighted sum r_{post}. Alongside the vectorized relation embeddings and the rhetorical encoder compute structural statistics that reflects discourse complexity that is demonstrated in Equation (5):

$$\begin{aligned} f_1 &= \frac{N_{\text{nucleus}}}{N_{\text{satellite}}} \\ f_2 &= \text{Average tree depth} \\ f_3 &= Entropy(p(r)), \quad p(r) = \frac{count(r)}{\sum_r count(r)} \end{aligned} \tag{5}$$

where f_1 denotes the nucleus-to-satellite ratio; here, N_{nucleus} and $N_{\text{satellite}}$ denote the total number of nucleus and satellite units, which is defined in the RST tree. A higher f_1 specifies the greater textual centrality and coherence, f_2 specifies the average tree depth for quantifying the hierarchical complexity of discourse structure. Moreover, f_3 specifies the rhetorical relations and $p(r)$ refers to the probability distribution of each relation type r. The tree depth is computed using term length, which is the longest path from root EDU to any leaf EDU that is defined in Equation (6):

$$Depth = \max_{e_k \in Leaves} d(e_{root},e_k) \tag{6}$$

where $d(e_{root},e_k)$ specifies the number of edges along with the path of the RST tree, and the relation entropy measures structural variability in the discourse and is defined using an empirical distribution of relation type in the tree, which is explained in Equation (7):

$$Entropy = -\sum_{r\in\mathcal{R}} p(r) \log p(r) \tag{7}$$

Here, $p(r)$ specifies the normalized frequency of relation type r among all edges in the RST tree. Further, the relation type is computed as the frequency r, which is divided by the total number of relations. A higher entropy value indicates more diverse rhetorical usage within the post. These features are concatenated with r_{post} weighted sum and passed through a projection layer to produce a final embedding that is depicted in Equation (8):

$$\mathrm{R} = \tanh(W_p[r_{\text{post}} \| f_1 \| f_2 \| f_3] + b_p) \tag{8}$$

Here, $\mathrm{R} \in \mathbb{R}^d$ specifies the final rhetorical embedding, which encodes how the thoughts in a post are connected. Here, W_p and b_p specify the learnable weights and bias; for instance, the depressive posts heavily rely on contrast or clause relations. Here, the NeuralRST parser is applied as a preprocessing step prior to the model training, and it is not the part of backpropagation pipeline. Each post is parsed once to extract the EDUs and discourse relations. Consequently, the rhetorical parsing process does not introduce additional gradient computation overhead.

3). BEHAVIORAL ENCODER. The behavioral encoder captures non-linguistic activity patterns that show how frequently users interact on social media. The behavioral encoder mainly concentrates on metadata-driven features such as posting frequency, participation patterns, and engagement level, which are often strong indicators for depression. Moreover, the temporal features such as posting frequency and inter-post intervals are long-scaled to reduce heavy-tailed variance and then standardized using Z-score normalization. The behavioral features are computed for each post from the considered attributes, which is depicted in Equation (9):

$$\begin{aligned} b_1 &= \Delta t_t = t_p - t_{p-1} \\ b_2 &= l_p \\ b_3 &= \frac{v_p + c_p}{\max(V,C)} \\ b_4 &= s_p \end{aligned} \tag{9}$$

where b_1 describes the time gap between consecutive posts t_p, b_2 specifies the normalized text length l_p, b_3 specifies the engagement ratio, and V and C specify the maximum number of upvotes and comments with the post, respectively. Moreover, $\max(V,C)$ specifies the normalized engagement values per user. These features collectively forms a behavioral vector as $b_t = [b_1,b_2,b_3,b_4] \in \mathbb{R}^d$. Consequently, an attention-based pooling mechanism is applied over time for aggregating the behavioral patterns, which is demonstrated in Equation (10):

$$\propto_t = \frac{\exp(w_b^T h_t^{(2)})}{\sum_{k=1}^{T} \exp(w_b^T h_t^{(2)})}, b_{\text{user}} = \sum_{t=1}^{T} \propto_t h_t^{(2)} \tag{10}$$

where $\propto_t$ denotes the significance of each post-behavioral context and the aggregated vector, $h_t^{(2)}$ denotes the output of the hidden layer, w_b^T specifies the learnable weight matrix of the behavioral cues, and b_{user} captures long-term tendencies such as irregular posting or engagement levels into dense embeddings, which represent user behavior that is associated with the depressive tendencies.

D. FEATURE FUSION WITH MAGM

The multimodal features are extracted separately through respective encoders; these embedding representations are combined into a unified embedding that reflects both semantic meaning and psychological context. Unlike the standard gating mechanisms, the MGAM differs mainly in two aspects. One is the gating coefficient, which is jointly conditioned on contextual, rhetorical, and behavioral embeddings rather than being independent per modality, allowing the cross-modal dependencies to influence weight allocation. Then, this attention mechanism integrates a relevance-aware suppression term, which down-weights the redundant discourses when the contextual semantics dominate. In contrast, it amplifies the rhetorical structures in posts exhibiting discourse irregularities. This dynamic cross-modal selection of gating attention mechanism enables the proposed model to learn modality-interaction patterns that are relevant to the specific post. This gating mechanism introduces an adaptive learnable gate to control the information flow from each encoder while ensuring balanced and context-sensitive integration.

The gating coefficients including g_S, g_R, and g_B are used for three modalities using a shared nonlinear transformation with Equation (11):

$$g_M = \sigma(W_g M + b_g) \tag{11}$$

Here, $M \in \{S,R,B\}$, $W_g \in \mathbb{R}^{d\times d}$, and $b_g \in \mathbb{R}^d$ are denoted as the learnable parameters, and $\sigma(\cdot)$ specifies the sigmoid activation function, which ensures that each gate value lies between 0 to 1. The gating vector g_M determines how much information from the particular modality needs to be retained for the current input post. For instance, if the user post contains rich linguistic signals but limited metadata, the contextual gate g_S holds a higher value than g_B. Consequently, each modality-gated output is computed with Equation (12):

$$M = g_M \odot \tag{12}$$

where $\odot$ denotes the element-wise multiplication, and this operation selectively suppresses features based on learned feature importance. These gated representations are concatenated and projected into a unified latent space, which is demonstrated in Equation (13):

$$F = \tanh(W_f[\tilde{S}\|\tilde{R}\|\tilde{B}] + b_f) \tag{13}$$

where $W_f \in \mathbb{R}^{d\times 3d}$ and $b_f \in \mathbb{R}^d$ specify learnable weight and bias parameters, respectively, the fused vector F represents a comprehensive multimodal feature embedding that balances text semantics, discourse structure, and behavioral activity. Moreover, $\tilde{S}$ specifies the gated semantic embedding, $\tilde{R}$ denotes the gated rhetorical embedding, and $\tilde{B}$ signifies the gated behavioral embedding. To ensure the convergence and stability, the adaptive gating mechanism is fully differentiable and trained end-to-end using backpropagation. The gating weights are computed using a SoftMax function over modality scores while ensuring that each weight lies in interval (0,1) that weights the sum to 1. This mechanism guarantees boundedness and forms a convex combination of modality embedding by preventing uncontrolled amplification during training. Since all the transformations are smooth and continuously differentiable, the gradients are well defined. The overall objective is optimized using the Adam optimizer with cross-entropy loss, which is a Lipschitz continuous with respect to network parameters. Under these conditions, the standard stochastic gradient-based optimization ensures stable convergence to local minimum. This empirical convergence behavior is further validated through consistent training and validation loss stabilization across epochs. The architecture of the MGAM attention is provided in Fig. 2.

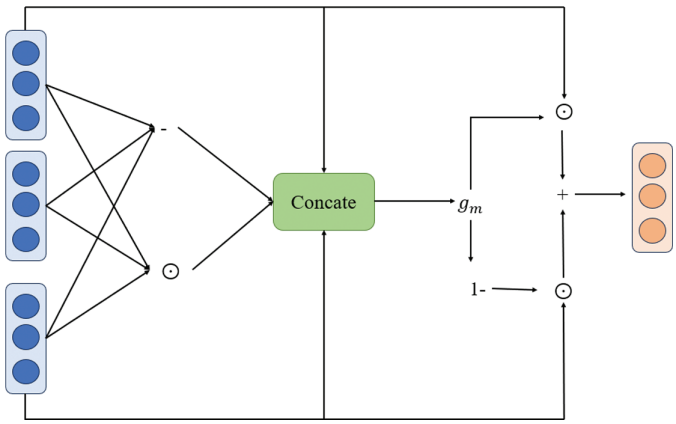

Fig. 2. The architecture diagram of the MGAM attention mechanism of the proposed DiscRST-HTT model for depression detection.

E. HIERARCHICAL TEMPORAL TRANSFORMER

The proposed DiscRST-HTT model explicitly couples the discourse structure with temporal evolution across user posts. Unlike the existing hierarchical model, which operates solely at document levels, the proposed DiscRST-HTT model integrates EDU-level discourse refinements during the intra-post stage and aligns them with the time-aware inter-post transitions using a learned bias matrix with posting intervals. This dual-granularity enables the HTT model to capture psychological progression patterns such as increasing discourse fragmentation or emotional drift, which are overlooked by the traditional hierarchical transformers. Therefore, this transformer model effectively provides a depression-specific temporal-discourse integration. Assume that a user u has a chronological sequence of T posts. After extraction and multimodal fusion, each post t is represented by a fused vector $F_t \in \mathbb{R}^d$, where F_t encodes the contextual, theoretical, and behavioral information. This HTT models temporal structure at two granularities is provided below.

1). INTRA-POST MODULE. When EDUs are available, then each post t has m_t EDUs with EDU-level contextual vectors $\{u_{t,1}, \ldots..,u_{t,m_t}\}$. This HTT first refines the F_t by attending to EDU dynamics with a cross-attention block as given in Equations (14) and (15):

$$Q_t = W_Q F_t, \quad K_{t,i} = W_K u_{t,i}, \quad V_{t,i} = W_V u_{t,i}, \tag{14}$$

$$Attn_{EDU}(F_t) = \sum_{i=1}^{m_t} \alpha_{t,i} V_{t,i}, \quad \alpha_{t,i} = \frac{\exp(Q_t^T K_{t,i}/\sqrt{d_k})}{\sum_j \exp(\ldots..)} \tag{15}$$

Here, Q_t, $K_{t,i}$, and $V_{t,i}$ refer a query, key, and value vectors at time t for a post i, and then the refined post summary is represented in Equation (16):

$$\tilde{F}_t = LayerNorm(F_t + FFN(Attn_{EDU}(F_t))) \tag{16}$$

This step explicitly fuses discourse-span signals with the fused vector by allowing the model to emphasize emotionally salient EDUs.

2). INTER-POST (USER-LEVEL) TEMPORAL TRANSFORMER: After the intra-post module, the user sequence is constructed as $\{\tilde{F}_1, \ldots \tilde{F}_T\}$. Then, to encode irregular posting intervals, HTT uses time-aware relative positional encodings. For posts i,j, it computes a scalar time gap $\Delta_{ij} = |t_i - t_j|$ and encodes it using a kernel $g(\Delta_{ij})$. When computing scaled dot-product attention, HTT injects a learned bias matrix B, which is dependent on Δ as given in Equation (17):

$$Attn(\tilde{F}) = softmax\left(\frac{QK^T}{\sqrt{d_k}} + \lambda B(\Delta)\right)V, \tag{17}$$

Here, $B(\Delta) = w_t^T \phi(\Delta)$ and $\phi(\cdot)$ specify the fixed bias such as log-time features, and this biases attention toward the temporal proximity and the relevant gaps in the user post. Additionally, the user-level transformer applies L layers on the attention output, which is given in Equation (18):

$$\begin{aligned} H^{(0)} &= \lfloor \tilde{F}_1; \ldots; \tilde{F}_T \rfloor, \\ H^{(\ell)} &= \text{Transformer Layer}_\ell(H^{(\ell-1)}, B(\Delta)) \end{aligned} \tag{18}$$

where $H^{(0)}$ specifies a transformer layer, and it produces the contextualized time-aware representations as z_t. A final attention pooling is applied to produce the user summary, which is given in Equation (19):

$$u = \sum_{t=1}^{T} \beta_t \mathbf{z}_t, \beta_t = \frac{\exp(\mathbf{v}^\top \tanh(W_z \mathbf{z}_t))}{\sum_k \exp(\ldots)} \tag{19}$$

Finally, an ordinal regression head is employed to learn ordered depression severity levels, from which the final binary predictions are derived through threshold-based mapping. Mathematically, the ordinal regression computes a score $s = w_t^T u + b$ from the final embedding of a user by representing the overall emotional intensity. In addition, the gradient flow between the ordinal regression head and temporal conditioning layers follows standard backpropagation. Here, h_t denotes the temporally conditioned representation and y denotes the ground-truth label. The regression head produces prediction $\hat{y} = W_r h_t + b_r$. The training loss L is computed using an ordinal regression objective. During optimization, gradients $\partial L/\partial h_t$ propagate from the regression head to the temporal conditioning module to enable joint parameter update across both components.

This layer employs a set of thresholds to model the ordered class boundaries. To ensure stable training, the thresholds are initialized with evenly spaced values within the output score range. This preserves natural ordering between sentiment classes. These parameters are subsequently optimized through gradient-based learning together with the network weights. The initialization strategy maintains the monotonic class separation while allowing the model to adaptively adjust decision boundaries on training data distribution. Moreover, the ordinal regression head uses learnable threshold parameters, which are initialized with the ordered values ($\tau_1 = -1.0$, $\tau_2 = 0.0$, $\tau_3 = 1.0$, $\tau_4 = 2.0$). These thresholds are optimized using backpropagation while maintaining the relative ordering. These values represent a boundary between classes including no depression, mild, moderate, and severe depression. The severity score is compared with the threshold, which is stated as the CORAL cumulative process, and it is mathematically given in Equation (20):

$$P(y > k|u) = \sigma(s - \tau_k) \tag{20}$$

Here, $P(y > k|u)$ defines the probability of a user post, σ specifies an activation function, s denotes the severity score, and τ_k specifies the threshold value. Based on the probability of threshold, the user is detected as depressed or not depressed, derived from ordinal severity levels. Unlike traditional binary classification models, the proposed framework learns ordered depression severity representations via ordinal regression before generating binary classification outputs. The integration of CORAL with the HTT model enforces monotonic severity progression and reduces label ambiguity frequency, which is observed in depression detection datasets. This usage of CORAL provides a stability and interpretability benefit as an output layer without temporal-discourse conditioning. Here, the model predicts ordered severity levels using a CORAL-based formulation ensuring rank consistency across classes. Here, the model predicts ordered severity levels using a CORAL-based formulation ensuring rank consistency across classes. However, the proposed model internally learns depression progression via ordinal severity prediction, and the final evaluation protocol follows the binary depression classification for consistency with the prior. The model learns ordered categories reflecting increasing levels of depression. The final binary classification is obtained by thresholding the learned ordinal severity predictions rather than training a separate binary classifier. Specifically, a user is classified as depressed if the predicted severity level exceeds a predefined threshold and non-depressed otherwise. For instance, the predicted ordinal label is

denoted with $\hat{y} \in \{0,1,2,3\}$ corresponding to nondepressed, mild, moderate, and severe. The binary prediction $\hat{y}_{\text{bin}}$ is derived as which is given in Equation (21):

$$\hat{y}_{\text{bin}} = \begin{cases} 0, & if\ \hat{y} = 0 \\ 1, & if\ \hat{y} \geq 1 \end{cases} \tag{21}$$

Therefore, this mapping preserves the ordinal structure while enabling direct comparison with prior classification methods.

IV. EXPERIMENTAL RESULTS

The proposed model is implemented on the simulation setup using PyTorch from Python 3.10 and HuggingFace transformer with NeuralRST for discourse parsing. The data are divided into 80:20 ratio for train and test. The model was trained on an NVIDIA A100 with 40 GB of memory using distributed multi-GPU training, and the AdamW optimizer was used with a learning rate of 1e-4, a weight decay of 1e-5, and a batch size of 8. Moreover, the maximum training epochs were set to 50; however, early stopping based on validation F1-score was applied, and the most experimental runs converged within 10 epochs. The selected epoch range on convergence stability is evaluated by monitoring the training and validation performance across epochs. The empirical observations indicate that the model converges relatively early with validation accuracy and stabilizing loss after the initial training phases. The increasing number of epochs beyond this point produced negligible performance improvement while increasing the risk of overfitting. Therefore, the chosen epoch range provides a practical balance between computational efficiency and stable model convergence. The evaluation metrics for the final binary classification derived from ordinal predictions, including accuracy, precision, recall, and F1-score as well as expected calibration error (ECE), are considered and mathematically expressed in Equations (22)–(27):

$$Accuracy = \frac{TP}{TP + FN + TN + FP} \tag{22}$$

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

$$Recall = \frac{TP}{TP + FN} \tag{24}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision \times Recall} \tag{25}$$

$$AUC = \int_a^b f(x)dx \tag{26}$$

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \begin{array}{l} accuracy(B_m) \\ -confidence(B_m) \end{array} \right| \tag{27}$$

Here, TP and TN denote the true positive and true negative, respectively, FP and FN signify the false positive and false negative, respectively, a and b specify the limits of function, and $f(x)$ specifies the function. In addition, N specifies the total number of data points and B_m denotes the group of samples whose predicted probability values fall into the interval $I_m = [((m-1)/M),(m/M)]$, and the perfectly calibrated models have an ECE of 0.

A. PERFORMANCE ANALYSIS

The performance of the proposed model is evaluated over baseline models including STCL, CAMF, and Temporal Modeling of Social Media for Depression Detection (TMSD) on the Reddit dataset under identical experimental conditions. All models are trained using the same optimizer, hyperparameters, and stratified data split to ensure fair comparison. Each baseline model represents a strong transformer-based architecture by incorporating semantic and sentiment features. The results of the proposed model over baseline models are given in Fig. 3.

From Fig. 3, the proposed model achieves better results across all the performance metrics with better results in accuracy and F1-score. The improvement in results is achieved from the ability to dynamically fuse heterogeneous features using the MAGM. This fusion mechanism enables better semantic, behavioral, and rhetorical modeling, while the HTT model effectively captures intra- and inter-post dependencies. Moreover, the performance comparison of sparse and low activity users for the dataset is illustrated in Fig. 4.

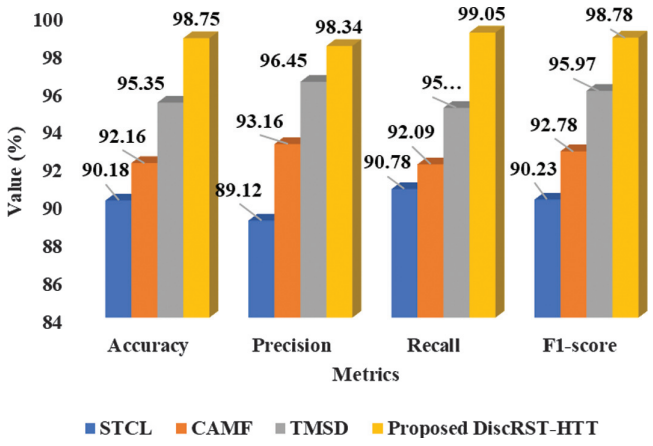

Fig. 3. Performance evaluation of binary depression classification derived from the ordinal severity predictions using the proposed DiscRST-HTT on Reddit dataset.

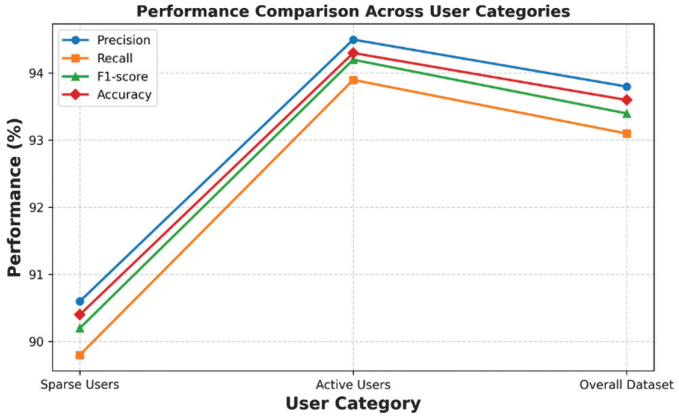

Fig. 4. Performance comparison of sparse users, active users, and the overall dataset.

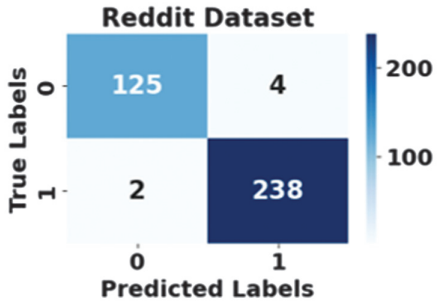

Fig. 5. The confusion matrix of the proposed DiscRST-HTT for Reddit dataset.

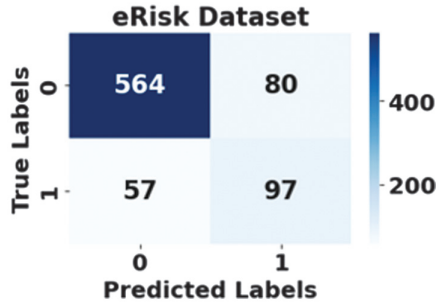

Fig. 6. The confusion matrix of the proposed DiscRST-HTT for eRisk dataset.

From Fig. 4, the proposed model achieves a consistent performance across various user activity levels. Although the scores for sparse user are slightly lower than for active users, the model still maintains strong predictive capability, indicating its robustness in low-activity behavioral scenarios. The confusion matrices of the proposed DiscRST-HTT for the Reddit and eRisk datasets are given in Figs. 5 and 6.

From Figs. 5 and 6, the confusion matrix diagonal elements exhibit consistently high values, which indicates that most samples are correctly classified within their respective categories. Very few misclassifications occur between visually similar posts from both datasets. Furthermore, to empirically justify the fixed maximum token length, which is used in the rhetorical encoder, a sensitivity analysis is conducted. It evaluates the model performance across multiple truncation thresholds, including 128, 256, and 384 tokens per EDU. The analysis is demonstrated in Table I.

From Fig. 1, the results indicate that the performance improves from 128 to 256 tokens but plateaus beyond 256, with negligible gains observed at 384 tokens. Additionally, the dataset statistics show that over 93% of EDUs fall within 256 tokens, indicating minimal truncation loss. Therefore, 256 tokens are selected as an optimal trade-off between computational efficiency and discourse coverage.

Table I. Sensitivity analysis for maximum token length in the rhetorical encoder

Token limit per EDU	Accuracy (%)	F1-score (%)
128 tokens	93.67	92.95
256 tokens	98.75	98.78
384 tokens	98.92	99.03

1). ADAPTIVE GATING MECHANISM EVALUATION. The MAGM is validated over traditional concatenation methods and repeated across five random seeds to minimize variance due to initialization and data shuffling. Each run maintains consistent hyperparameters and is trained for 10 epochs with early stopping on validation performance. The results are reported with the mean $\pm$ standard deviation of the performance metrics across runs. The result of this evaluation is given in Table II.

From Table II, the gating mechanism in the proposed model clearly outperforms the traditional fusion models by achieving the highest accuracy with the lowest ECE value. Unlike the traditional fusion models, which combine modalities with uniform weighting, the proposed gating mechanism dynamically learns the importance of each modality based on contextual relevance. This adaptive weighting enhances the information integration while ensuring better calibration of confidence scores, resulting in more reliable depression detection outcomes.

2). GENERALIZATION THROUGH CROSS-DATASET VALIDATION. The generalization capability of the proposed model is evaluated over three baseline models including STCL, CAMF, and TMSD on the unseen Dereddit dataset [28]. Each model is trained and evaluated under identical experimental configurations, which are repeated across five random seeds and five independent runs to ensure statistical consistency. The results are reported with mean $\pm$ and standard deviation to confirm robust comparative analysis. This evaluation ensures the cross-domain generalization and reliability of multimodal temporal representations beyond the original training dataset, which is provided in Table III.

From Table III, the proposed model achieves the highest mean performance with the lowest variance across all the metrics, thereby demonstrating the strong generalization to unseen data. This improved performance is achieved through the integration of multimodal encoders via an adaptive gating mechanism. Additionally, the HTT model effectively models the user-specific discourse evolution by allowing the consistent depression detection even with unseen linguistic variations. Moreover, this generalization ability is evaluated on the datasets, which are derived from platforms with relatively similar structural and interaction characteristics. While this evaluation allows for consistent comparison and controlled experimentation, it may limit the generalizability of the model to social media environments with different user behaviors or interaction patterns.

3). ABLATION AND STATISTICAL ANALYSIS. The ablation study and statistical analysis of the proposed model are evaluated across five random seeds with three independent runs per configuration to ensure reproducibility. Excluding each component in the proposed model assesses its contribution under identical experimental settings. The results are reported as the mean accuracy with the standard deviation including 95% confidence interval (CI) and *p*-test significance. The results are given in Table IV.

Table II. The evaluation of a multimodal adaptive gating mechanism over traditional fusion models

Fusion models	Accuracy (%)	ECE
Additive fusion	94.62 ± 1.5	0.072
Multiplicative fusion	95.18 ± 1.2	0.065
Concatenation	96.03 ± 0.8	0.058
Gating mechanism in proposed DiscRST-HTT	98.75 ± 0.5	0.031

Table III. Evaluating the generalizability of the proposed DiscRST-HTT model with the unseen Dereddit dataset

Performance model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
STCL	85.10 ± 0.42	86.28 ± 0.37	85.17 ± 0.45	85.78 ± 0.41
CAMF	88.35 ± 0.38	86.15 ± 0.33	87.59 ± 0.40	86.89 ± 0.35
TMSD	90.16 ± 0.36	88.18 ± 0.31	89.48 ± 0.34	88.94 ± 0.33
Proposed DiscRST-HTT	95.18 ± 0.28	94.89 ± 0.26	95.35 ± 0.29	95.04 ± 0.27

Table IV. Evaluating the effectiveness of individual components in the proposed DiscRST-HTT with statistical significance

Model variants	Accuracy (%)	Mean ± std	95% CI	*p*-Test
Without multimodal encoder	94.21	94.21 ± 0.47	[93.72, 94.69]	0.0041
Multimodal encoder (without Z-score normalization)	94.89	94.89 ± 0.40	[94.50, 95.10]	0.0038
Multimodal encoder (4 heads + Z-score normalization)	95.12	95.12 ± 0.38	[94.90, 95.30]	0.0036
Multimodal encoder (8 heads)	95.32	95.32 ± 0.36	[94.95, 95.40]	0.0034
Multimodal encoder (12 heads)	95.22	95.22 ± 0.37	[94.90, 95.30]	0.0037
Without gating mechanism	95.63	95.63 ± 0.39	[95.20, 96.05]	0.0032
Without preprocessing and ordinal severity	96.12	96.12 ± 0.35	[95.75, 96.49]	0.0029
Proposed DiscRST-HTT	98.75	98.85 ± 0.28	[98.52, 99.13]	0.0015

From Table IV, the proposed model achieves the highest mean accuracy with a narrow CI, which significantly outperforms all the variants with smallest p-value (<0.005). The superior performance indicates that the integration of a multimodal encoder, adaptive gating mechanism, and preprocessing significantly enhances the generalization. The removal of Z-score normalization from the multimodal encoder reduces the slight accuracy of the proposed model. Moreover, the robustness of the attention mechanism is assessed by conducting a sensitivity analysis on the key attention-related hyperparameters within the encoder. The normalization is used mainly to standardize variations across multiple impacted users by ensuring that user-specific differences do not bias the learned representations. Experiments are performed by varying configurations such as the number of attention heads and attention weight scaling parameters. The results indicate moderate changes in these parameters, which produce minor variations in the model performance. Therefore, for analyzing the encoder architecture remains stable across a range of attention settings. Moreover, to ensure scalability, the proposed model is supported by modular transformer and batch-based training strategy. The encoder structure is parallelized across GPUs due to independent working before fusion, which enables efficient processing of large-scale datasets. In addition, the HTT processes user sequences with bounded token lengths while ensuring computational complexity grows approximately linearly with the number of posts. This design allows the model to scale effectively to large social media datasets without substantial increases in processing time. Moreover, a runtime scaling experiment is conducted by progressively increasing the dataset size and measuring the corresponding training time. It ensures computational scalability of the proposed framework while indicating predictable processing overhead under expanded training conditions. The low standard deviation across runs confirms the training stability, while the statistical significance validates the effectiveness of each component for depression detection.

4). COMPUTATIONAL ANALYSIS. The computational efficiency of the proposed DiscRST-HTT model is evaluated with its ablated variants. Metrics such as execution time, inference latency, parameter count, and memory utilization with FLOPs are considered. This analysis is performed to validate the model's scalability, efficiency, and trade-off between accuracy and complexity, which is presented in Table V.

From Table V, the computational analysis of the proposed model demonstrates lower inference latency and optimized parameter utilization compared to its ablation variants. This result indicates that the integration of multimodal encoders and adaptive gating enhances representational ability while maintaining computational efficiency.

B. COMPARATIVE ANALYSIS

The proposed model is compared with existing models on the Reddit dataset, which is presented in Table VI. The existing models, including RSTFusionX [19], N-gram-based Transformer [21], DLAD [22], EHL [23], and BERT + CNN [24] are

Table V. Evaluating the computational complexity of the proposed model variants for depression detection

Model variants	Parameters (M)	FLOPs (G)	Execution time (s/epoch)	Inference time (ms/sample)	GPU memory (GB)
Without multimodal encoder	78.5	18.2	54.6	42.8	9.6
Without gating mechanism	82.3	19.7	58.9	46.3	10.1
Without preprocessing	84.7	20.1	60.2	48.5	10.3
Proposed DiscRST-HTT	86.9	21.4	62.8	49.2	10.7

Table VI. Comparative analysis of proposed DiscRST-HTT over reported baselines from the literature review on the Reddit dataset

Comparative models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
RSTFusionX [19]	97.10	97.40	96.70	96.90
N-gram-based transformer [21]	N/A	91.00	88.00	89.00
DLAD [22]	96.00	94.00	98.00	95.96
EHL [23]	75.12	81.15	75.12	77.01
BERT + CNN [24]	90.50	92.40	84.30	89.70
Proposed DiscRST-HTT	98.75	98.34	99.05	98.78

considered. The evaluation is done with performance metrics such as accuracy, precision, recall, and F1-score as shown in Table VI.

From Table VI, the proposed model attains a higher accuracy and F1-score, which demonstrates the effectiveness of combining semantic, discourse, and behavioral cues with adaptive weighting. The existing models attain fewer results than the proposed model, as they mainly work based on single-modality and simple concatenation models. The proposed model dynamically highlights the relevant modalities for each post while ensuring improved interpretability and reduced overfitting. The proposed model attains better results in terms of accuracy (98.75%), precision (98.34%), recall (99.05%), and F1-score (98.78%) on the Reddit dataset. The HTT effectively captured the temporal evolution of depressive severity patterns, which improved final binary depression classification performance. Moreover, the proposed model is compared with the existing EHL [23] model to ensure generalizability on the eRisk dataset. The comparative results are given in Table VII.

From Table VII, the proposed model achieves superior results by balancing contextual semantics and behavioral indicators through dynamic gating, while HTT captures longitudinal depression cues. The proposed model enhances sensitivity to subtle emotional and discourse transitions while leading to more reliable depression detection. The proposed model attains strong results including accuracy (82.56%), precision (81.07%), recall (83.48%), and F1-score (82.39%). The AUC–ROC curves of the proposed DiscRST-HTT model for both datasets are provided in Figs. 7 and 8.

Figures 7 and 8 demonstrate the AUC–ROC performance of the proposed model to distinguish between depression classes.

Each class exhibits an AUC value close to 1.0, which indicates separability between different classes. The rise in ROC space demonstrates a high true-positive rate with minimal false positives across all classes. This performance highlights the strong discriminative power, which is enhanced by multi-scale feature extraction and attention refinement to enable accurate detection of subtle and overlapping regions. Moreover, the training and validation curve for the proposed model on the Reddit dataset is illustrated in Fig. 9.

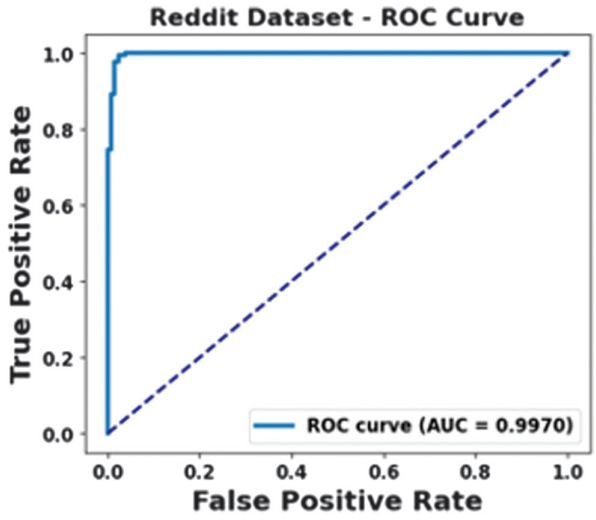

Fig. 7. AUC–ROC performance of the proposed DiscRST-HTT model on Reddit dataset.

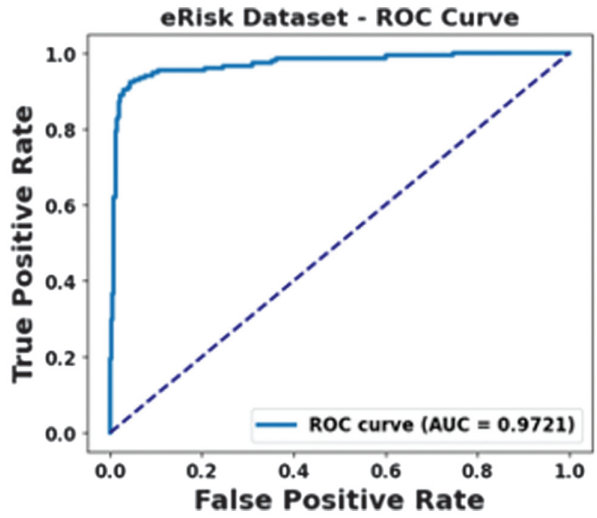

Fig. 8. AUC–ROC performance of the proposed DiscRST-HTT model on eRisk dataset.

Table VII. Comparative analysis of the proposed DiscRST-HTT over existing models on eRisk dataset

Comparative models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EHL [23]	75.55	80.05	74.55	76.55
Proposed DiscRST-HTT	82.56	81.07	83.48	82.39

From Fig. 9, the loss curve demonstrates the stable convergence of the proposed model. The loss decreases rapidly during the early epochs and gradually stabilized as training progresses. The close alignment between the curves indicates the effective learning of the proposed model without significant overfitting.

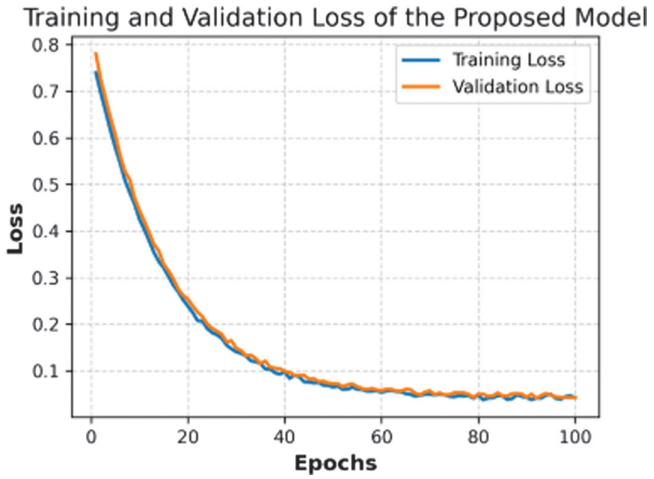

Fig. 9. The training and validation loss curve for the proposed DiscRST-HTT model on Reddit dataset.

V. RESEARCH IMPLICATIONS

The proposed DiscRST-HTT model introduces various research implications that advance the field of computational mental health analysis. Initially, the combination of rhetorical structure theory with the contextual transformer embeddings generates a novel linguistic-cognitive perspective for a deeper understanding of how the thought organization and emotional coherence reflect depressive expression. Moreover, the proposed model addresses the label biasing issue with the stratified dataset splitting to maintain consistent data distributions across training, validation, and testing sets. Additionally, text preprocessing and normalization steps reduce annotation inconsistencies and platform artifacts. Further, the evaluation of the model across multiple datasets further minimizes dataset-specific bias and helps to ensure learned representations. In addition, the MAGM exceeds the static concatenation approaches for ensuring interpretability and balanced contribution across textual, discourse, and behavioral modalities, respectively. Also, the inclusion of behavioral cues establishes the importance of non-linguistic user activity patterns in complementing the textual semantics for early and reliable detection. To further interpret the model's capability in psychological patterns, the posts containing strong emotional expressions such as frustration or self-doubt, which are often associated with the specific rhetorical structures including justification relations, are considered. Subsequently, informational posts exhibit simpler discourse structures and lower emotional intensity. For instance, the posts such as "I feel exhausted but I am trying to stay Positive" contain mixed emotional cues, which sometimes lead the model to classify them as nondepressed class despite underlying distress. These examples illustrate how linguistic cues, discourse structure, and activity patterns collectively reflect underlying psychological tendencies. Moreover, the HTT presents a dual-level temporal modeling paradigm to capture both intra-post discourse variations and inter-post emotional transitions effectively for improving the sensitivity to evolving depressive signals. Furthermore, the calibrated ordinal regression head strengthens the model's representational capability by learning fine-grained ordered depression severity patterns that support reliable binary depression classification. Here, the ordinal formulation enables the model to learn progressive depression severity representations before converting them into binary predictions for standardized evaluation. As part of experimental validation, a qualitative error analysis is conducted on misclassified samples. Several samples are identified where the posts exhibit higher similar linguistic expressions and rhetorical structures across various classes. For example, for the posts such as "I feel like nothing I do matters anyone," it poses self-doubt with justification. In such cases, the subtle contextual cues make the class boundaries complex to distinguish. Finally, these contributions improve predictive performance and trustworthy AI systems for responsible deployment of multimodal transformer architectures in mental health research and personalized digital well-being monitoring.

VI. CONCLUSION

In this research, a novel DiscRST-HierAT framework is presented for the detection of social media-based depression by integrating contextual, rhetorical, and behavioral representations in a unified transformer architecture. Here, the contextual semantics are extracted by using MentalBERT, and then the NeuralRST captures discourse coherence through rhetorical relations, and finally, the behavioral cues quantify user engagement and temporal activity. Further, these multimodal features are adaptively fused by using a MAGM, which allows dynamic weighting of modality contributions. After that, the fused embeddings are processed by an HTT, which models the intra-post and inter-post dependencies. Finally, an ordinal regression head with calibration is employed to learn ordered depression severity representations, which are subsequently transformed into binary depression predictions for evaluation and benchmarking. The methodology provides improved robustness and sensitivity to linguistic and behavioral progression by learning structured ordinal severity representations that enhance final binary depression classification performance. In the future, this research will focus on incorporating multilingual discourse modeling by expanding to multimodal signals and exploiting self-supervised pretraining to improve the cross-domain generalization and real-world clinical applicability.

CONFLICT OF INTEREST STATEMENT

The author(s) declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] S. Zhou and M. Mohd, "Mental health safety and depression detection in social media text data: A classification approach based on a deep learning model," *IEEE Access*, vol. 13, pp. 63284–63297, 2025.

[2] R. Chionga, G. S. Budhib, and E. Cambriac, "Detecting signs of depression using social media texts through an ensemble of ensemble classifiers," *IEEE Trans. Affect. Comput.*, vol. 16, no. 4, pp. 2931–2945, 2025.

[3] Z. Ding *et al.*, "Trade-offs between machine learning and deep learning for mental illness detection on social media," *Sci. Rep.*, vol. 15, no. 1, p. 14497, 2025.

[4] G. A. Pradnyana *et al.*, "Revealing depression through social media via adaptive gated cross-modal fusion augmented with insights from personality traits," *IEEE Access*, vol. 13, pp. 133465–133482, 2025.

[5] D. Suganthi and A. Geetha, "Predicting postpartum depression with aid of social media texts using optimized machine learning model," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 3, pp. 417–427, 2024.
[6] S. Velu and P. Muthuswamy, "Multi level learning based hierarchical convolutional neural network-long short-term memory model for depression intensity prediction," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 4, pp. 1122–1133, 2024.
[7] Z. Guo *et al.*, "Leveraging domain knowledge to improve depression detection on Chinese social media," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 4, pp. 1528–1536, 2023.
[8] M. Jain *et al.*, "CDME-GAT: Context-aware depression detection using multiembedding and graph attention networks in social media text," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 6, pp. 7212–7222, 2024.
[9] A. Anshul *et al.*, "A multimodal framework for depression detection during covid-19 via harvesting social media," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 2, pp. 2872–2888, 2023.
[10] M. A. Abbas *et al.*, "Novel transformer based contextualized embedding and probabilistic features for depression detection from social media," *IEEE Access*, vol. 12, pp. 54087–54100, 2024.
[11] H. Zogan *et al.*, "Hierarchical convolutional attention network for depression detection on social media and its impact during pandemic," *IEEE J. Biomed. Health. Inform.*, vol. 28, no. 4, pp. 1815–1823, 2023.
[12] S. A. Asma *et al.*, "Hierarchical explainable network for investigating depression from multilingual textual data," *IEEE Access*, vol. 12, pp. 131915–131927, 2024.
[13] S. Almutairi *et al.*, "A hybrid deep learning model for predicting depression symptoms from large-scale textual dataset," *IEEE Access*, vol. 12, pp. 168477–168499, 2024.
[14] D. S. Nair and M. Lalithambigai, "ACSAT: An adaptive contextual sentiment-aware transformer for early detection of depression in social media text," *Indian J. Sci. Technol.*, vol. 18, no. 34, pp. 2752–2760, 2025.
[15] M. Kerasiotis, L. Ilias, and D. Askounis, "Depression detection in social media posts using transformer-based models and auxiliary features," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, p. 196, 2024.
[16] D. Agarwal *et al.*, "Dwarf updated pelican optimization algorithm for depression and suicide detection from social media," *Psychiatr Q.*, vol. 96, pp. 529–562, 2025.
[17] W. Zhang, K. Mao, and J. Chen, "A multimodal approach for detection and assessment of depression using text, audio and video," *Phenomics*, vol. 4, no. 3, pp. 234–249, 2024.
[18] H. Hamidi and M. Ghelichi, "A hybrid deep learning technique for depression detection of Persian tweets using using Bi-directional LSTM-CNN model," *Multimed. Tools Appl.*, vol. 84, pp. 48779–48809, 2025. DOI: 10.1007/s11042-025-21092-7.
[19] S. Ajmal, M. Shoaib, and F. Iqbal, "RSTFusionX: Leveraging rhetorical structure theory and ensemble models for depression prediction in social media posts," *IEEE Access*, vol. 12, pp. 118389–118404, 2024.
[20] R. Chiong *et al.*, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Comput. Biol. Med.*, vol. 135, p. 104499, 2021.
[21] A. Qasim *et al.*, "Detection of depression severity in social media text using transformer-based models," *Information*, vol. 16, no. 2, p. 114, 2025.
[22] K. Salameh *et al.*, "Deep linguistic analysis for depression in social media using RoBERTa and CNN," *Int. J. Speech Technol.*, vol. 28, no. 4, pp. 825–836, 2025.
[23] L. Ansari *et al.*, "Ensemble hybrid learning methods for automated depression detection," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 1, pp. 211–219, 2022.
[24] C. Xin and L. Q. Zakaria, "Integrating BERT with CNN and BiLSTM for explainable detection of depression in social media contents," *IEEE Access*, vol. 12, pp. 161203–161212, 2024.
[25] L. Ilias, S. Mouzakitis, and D. Askounis, "Calibration of transformer-based models for identifying stress and depression in social media," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 2, pp. 1979–1990, 2023.
[26] Reddit Dataset: https://www.kaggle.com/datasets/rishabhkausish/reddit-depression-dataset (Accessed on September 2025).
[27] eRisk Dataset: https://www.kaggle.com/datasets/albertouah/erisk-feature-engine-v2/data (Accessed on September 2025).
[28] E. Turcan and K. McKeown, "Dreaddit: A reddit dataset for stress analysis in social media", arXiv preprint arXiv:1911.00133, 2019.