Received 16 June 2025, accepted 21 July 2025, date of publication 28 July 2025, date of current version 1 August 2025.

*Digital Object Identifier 10.1109/ACCESS.2025.3593273*

RESEARCH ARTICLE

# Revealing Depression Through Social Media via Adaptive Gated Cross-Modal Fusion Augmented With Insights From Personality Traits

**GEDE ADITRA PRADNYANA**[1], **(Graduate Student Member, IEEE), WIWIK ANGGRAENI**[2], **(Member, IEEE), EKO MULYANTO YUNIARNO**[1,3], **(Member, IEEE), AND MAURIDHI HERY PURNOMO**[1,3,4], **(Senior Member, IEEE)**

[1]Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
[2]Department of Information Systems, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
[3]Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
[4]University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS), Surabaya 60111, Indonesia

Corresponding author: Mauridhi Hery Purnomo (hery@ee.its.ac.id)

This work was supported by Indonesian Education Scholarship (BPI), through the Center for Higher Education Funding and Assessment (PPAPT) of the Ministry of Higher Education, Science, and Technology, Indonesia; and by Indonesian Endowment Fund for Education (LPDP).

**ABSTRACT** The escalating prevalence of mental health disorders, particularly depression, underscores the urgent need for early and accurate detection systems. The widespread use of social media has produced an extensive repository of user-generated content, offering critical insight into individuals' emotional and psychological states. This development enables novel approaches for automated depression detection. However, existing multimodal depression detection approaches often adopt rigid fusion strategies and disregard individual differences in expressive behavior by adopting generalized, one-size-fits-all frameworks. To bridge this gap, we introduce DeXMAG, a novel personalized depression detection framework that integrates a Cross-Modal Attention mechanism with an Adaptive Gated Fusion strategy. DeXMAG effectively combines textual content, visual imagery, and user-level personality traits to capture the heterogeneous and individualized patterns of depressive expression across modalities. Leveraging domain-specific transformer-based language models and a convolutional neural networks-based model, we extract modality-specific features and fuse them via a novel adaptive gating mechanism. Unlike traditional fusion techniques, our model leverages modality-aware attention signals and gating coefficients to enable flexible and personalized integration of features. An extensive ablation study reveals that the most substantial performance gain occurs when DeXMAG is augmented with insights from Myers–Briggs Type Indicator (MBTI) personality traits in conjunction with textual and visual features. Comprehensive experiments on multimodal social media datasets demonstrate that DeXMAG consistently outperforms unimodal baselines and state-of-the-art approaches, achieving an accuracy of 95.20% and an F1-score of 95.94%. These results reinforce the effectiveness of adaptive gated fusion and personality-aware modeling in advancing robust and personalized depression detection.

**INDEX TERMS** Adaptive gated fusion, cross-modal attention, depression, multimodal learning, Myers–Briggs Type Indicator, personalized detection, social media.

## I. INTRODUCTION

Depression, a prevalent and debilitating mental health disorder, affects over 280 million people worldwide and contributes significantly to the global burden of disease

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

according to the World Health Organization [1], [2], [3], [4]. Depression is a complex and multifactorial mental health disorder characterized by prolonged feelings of sadness, diminished interest or pleasure in daily activities, and a range of cognitive and somatic symptoms. Beyond impairing daily functioning, depression is also a major contributor to the global rise in suicide rates [5], [6]. Traditionally, depression diagnosis relies on clinical interviews and standardized self-reported questionnaires such as PHQ-8 or PHQ-9, which are often time-consuming, subjective, and inaccessible to individuals reluctant to seek help due to stigma or limited healthcare access [5], [7]. These limitations have catalyzed a paradigm shift towards developing automatic, AI-based depression detection systems that facilitate early intervention, scalability, and affordability.

The rapid evolution of Artificial Intelligence (AI) has enabled the automatic inference of mental states through text, audio, and visual data using both traditional machine learning (ML) and, more recently, transformer-based deep learning models [2], [8], [9]. Transformer architectures such as BERT [10]] and RoBERTa [11] have demonstrated superior performance in depression classification due to their ability to capture contextual semantics in large-scale language representations [2], [9], [12]. Several studies have enhanced these models with domain-specific tuning and hybrid learning techniques, showcasing their efficacy across benchmark datasets [13].

Among the diverse data sources used for AI-driven mental health assessment, social media has emerged as a rich and promising medium [8], [14], [15]. With the rapid advancement of internet and information technologies, social media has become a perceived safe space where many individuals who are concerned about stigma or personal discomfort in seeking professional help choose to express emotional distress. Such expressions are often conveyed indirectly, either through pseudonymous accounts or via emotionally valenced posts that respond to daily events, media content, or social interactions. Prior research has shown that this type of indirect or metaphorical expression, which may include frustration, hopelessness, or affective reflections, can provide meaningful indicators of underlying psychological states [4], [9], [16], [17]. A recent surveys indicate that individuals aged 14 to 22 increasingly use these platforms to share their emotional experiences on a near-daily basis [4]. Social media platforms like Twitter, Reddit, and Facebook host massive volumes of user-generated content that reflect individuals' emotions, behaviours, and social interactions. This content provides unobtrusive and real-time signals that can be leveraged to infer depressive symptoms at scale. In particular, studies have shown that users often share emotionally charged language and depressive expressions more openly in online platforms than in traditional clinical settings [1], [18].

Given the inherently multimodal nature of content shared on social media, including textual posts, images, video recordings, and behavioral interaction patterns, relying solely on a single modality may constrain the ability to comprehensively detect depressive cues [19]. Despite the promise of unimodal approaches, recent studies underscore the importance of multimodal fusion for enhancing detection accuracy [1], [20], [21], [22], [23], [24], [25]. Depression manifests not only through linguistic patterns, but also through acoustic (e.g., tone, pitch) [6], [26], visual (e.g., facial expressions) [27] and behavioral signals (e.g., posting frequency or late-night post) [4], [18]. Hence, several studies have adopted multimodal learning paradigms, combining textual, acoustic, and visual features to achieve robust mental state inference. The multimodal approach has demonstrated consistent performance improvements in both clinical and social media-based datasets [1], [3], [20], [24], [28], [29].

Static and attention-based cross-modal fusion have emerged as two prominent methodologies in multimodal depression detection research. Static fusion, also commonly categorized as early or late fusion, represents one of the most straightforward strategies for integrating multimodal data [30] In early fusion, low-level features extracted from different modalities (e.g., text and image) are concatenated or combined using simple operations such as summation or averaging before being processed by a unified model. In contrast, late fusion involves the aggregation of high-level predictions, such as class probabilities or logits, produced independently by unimodal classifiers. While static fusion is computationally efficient and relatively simple to implement, it is inherently limited by its inability to dynamically modulate the relative importance of each modality [4]. This limitation is particularly problematic in real-world social media data, where the informativeness of each modality can vary substantially across users and instances. For example, certain users may predominantly express depressive symptoms through linguistic cues, while others may do so through visual content.

To address the rigidity and lack of adaptivity in static fusion, attention-based cross-modal fusion mechanisms have been proposed [4], [6], [18], [31], [32]. These mechanisms have gained increasing traction due to their capability to align semantically meaningful representations across disparate modalities, thereby enabling more nuanced and context-aware integration. Notable frameworks such as the Additive Cross-Modal Attention Network [6], Cross-Modal Attention Transformers [18], and Hierarchical Attention Networks [33] have consistently demonstrated superior performance over traditional baselines by capturing fine-grained inter-modal dependencies and leveraging modality interactions that are sensitive to both context and task-specific signals.

However, despite their flexibility, attention-based fusion models exhibit specific vulnerabilities. They tend to perform suboptimally when one modality is noisy, sparse, or semantically weak, a common characteristic in user-generated social media content [19]. While static and attention-based approaches contribute valuable perspectives to multimodal learning, neither fully addresses the challenge of modality reliability heterogeneity. Moreover, attention mechanisms alone do not inherently suppress irrelevant or misleading modality signals. These mechanisms merely reweight features based on local interactions, which may still propagate noise if not regulated. To overcome these limitations, we propose a cross-modal fusion framework with adaptive gated fusion, which explicitly learns gating functions to modulate the flow of modality-specific features based on contextual and personalized cues. These gates, parameterized via neural networks with sigmoid activations, act as soft filters that selectively enhance or suppress modality contributions at the feature level.

Another significant limitation of current multimodal depression detection systems is their lack of personalization. Most existing models adopt a generalized, one-size-fits-all framework, implicitly assuming that depressive symptoms manifest uniformly across all users. On the other hand, a substantial body of psychological research underscores that individual differences, particularly personality traits, influence how emotional states are internally experienced and externally expressed [34], [35]. For example, a social media post stating "*I feel very alone and exhausted . . .*" may reflect different psychological implications depending on whether the author is an extrovert or an introvert. Empirical studies in psychology have consistently demonstrated that specific personality dimensions are significantly correlated with increased vulnerability to depressive symptoms [36], [37], [38]. Conversely, traits such as extraversion have been linked to lower levels of depression and an overall improved quality of life [35], [38]. Individuals exhibiting low levels of conscientiousness and extraversion, combined with psychological attributes such as self-critical thinking, dependency, and fixation, tend to present moderate to severe depressive tendencies. These findings suggest that individual personality profiles can serve as important predictive factors in assessing mental health risk, particularly in the context of depression.

In light of these insights, our study leverages personality information to improve the recognition of depressive symptoms and the likelihood of depression onset. Psychological findings support the hypothesis that incorporating user-level characteristics can reduce the risk of misinterpretation and improve the model's capacity to detect at-risk individuals more accurately. Accordingly, we posit that integrating user-specific features, such as personality traits, into AI-based models enables adaptive reasoning that accounts for the heterogeneity in users' expressive styles, thereby enhancing the robustness and personalization of multimodal depression detection systems. To operationalize this, we incorporate the Myers-Briggs Type Indicator (MBTI) framework, a widely used personality typology that classifies individuals across four dichotomous dimensions: Introversion(I)-Extraversion(E), Sensing(S)-Intuition(N), Thinking(T)-Feeling(F), and Judging(J)–Perceiving(P). Prior research has demonstrated significant associations between certain MBTI types and higher susceptibility to depression [36], [39]. By embedding MBTI-based personality profiles as explicit input features in our detection model, we aim to enable more nuanced interpretations of linguistic and behavioral cues, particularly those that might be overlooked in generic models. This integration supports a more personalized and psychologically informed approach to depression detection, allowing the system to better account for individual variation in emotional expression and vulnerability.

In this study, we propose DeXMAG (Depression Detection via Cross-Modal Attention and Adaptive Gated Fusion), a novel personalized multimodal deep learning framework designed to enhance depression detection from social media. The proposed framework integrates multiple modalities, including textual content and visual imagery, while incorporating user-specific psychological attributes to achieve a more context-aware and personalized mental health assessment. The key contributions of DeXMAG are summarized as follows:

- To enable personalized depression detection, our DeXMAG integrates personality information as MBTI-based personality traits. To the best of our knowledge, this is the first study to explore the integration of personality traits into a multimodal depression detection framework. The inclusion of personality features is motivated by a substantial body of psychological research demonstrating that personality traits are strongly associated with the onset, severity, and progression of mental health disorders.
- Our proposed DeXMAG model incorporates a Cross-Modal Attention (CMA) mechanism to facilitate deep interaction between features derived from text and image modalities. The feature representations are extracted using domain-specific encoders tailored to each modality. We utilize general-purpose and domain-specific transformer-based language models, including RoBERTa and MentalRoBERTa, to extract rich contextual representations from textual data. We use convolutional neural network (CNN) architectures for image data, specifically VGG16 and ResNet50, to extract hierarchical visual features. The CMA module aligns and fuses these multimodal representations by enabling mutual attention between textual and visual features, thereby capturing complementary information critical for personalized depression detection.
- In contrast to previous studies that rely on static feature concatenation, the proposed DeXMAG model introduces an Adaptive Gated Fusion (AGF) mechanism that dynamically modulates the contribution of each modality based on its contextual relevance. Unlike static fusion, which treats all modalities equally informative regardless of the input instance, the adaptive gating mechanism learns modality-specific weights conditioned on the input features. This dynamic modulation

allows DeXMAG to selectively emphasize or attenuate modality-specific signals depending on the user's data characteristics. As a result, this AGF enhances the model's robustness to modality imbalance and noise, while improving generalization to diverse user behaviors and content styles commonly observed in social media environments.

By leveraging cross-modal attention, adaptive gated fusion, and personalized features, DeXMAG provides a robust solution for AI-driven mental health assessment, addressing the key challenges in multimodal depression detection from social media. This paper provides a comprehensive explanation of the proposed DeXMAG model and presents the research we conducted on its effectiveness for multimodal depression detection. The remainder of the paper is organized as follows: Section II reviews related work in the domains of multimodal depression detection, gated fusion strategies, and personality-aware AI models. Section III introduces the proposed DeXMAG architecture, elaborating on the cross-modal attention mechanism and the adaptive gated fusion module. Section IV presents the experimental framework, encompassing dataset descriptions, feature modalities, evaluation metrics, and a detailed analysis of the results. This includes a comparative performance assessment against existing state-of-the-art models. Finally, Section V concludes the paper and outlines potential directions for future research.

## II. RELATED WORKS

### A. TOWARD MULTIMODAL APPROACHES IN DEPRESSION DETECTION

In recent years, significant progress has been made in developing automatic depression detection systems that leverage social media data [8], [14]. These advancements span from unimodal approaches, which rely on a single modality such as text or image, to more sophisticated multimodal frameworks that integrate heterogeneous data sources. This evolution reflects a growing recognition of the multifaceted nature of depressive behaviors as expressed in online environments.

Early studies primarily focused on text-based classification tasks, utilizing user-generated content from Twitter, Reddit, and Facebook platforms. These approaches harness linguistic features, ranging from lexical choices to syntactic patterns, and process them using classical machine learning classifiers or deep learning architectures. Several studies have demonstrated the effectiveness of deep learning models for text-based depression detection. For instance, Thekkekara et al. [40] proposed a hybrid architecture that combines a Convolutional Neural Network (CNN) and a Bidirectional Long Short-Term Memory (BiLSTM) network, enhanced with an attention mechanism to extract salient features from user-generated text for depression classification. More recently, transformer-based models such as BERT, RoBERTa, and DistilBERT have been widely adopted due to their superior ability to capture contextual dependencies in language. Kerasiotis et al. [41] demonstrated the effectiveness of DistilBERT augmented with metadata for depression detection, achieving an F1-score exceeding 84%. Similarly, Poświata and Perełkiewicz [42] fine-tuned RoBERTa for depression classification in the LT-EDI-ACL2022 shared task and reported promising macro-averaged F1-scores. In parallel, image-only approaches have explored visual cues such as color tone, facial expression, and aesthetic composition in social media posts. Nazira et al. [43] developed a facial expression recognition system for depression detection, employing a combination of CNN, OpenCV, and the Haar Cascade Classifier to extract and analyze facial features indicative of depressive states. However, image-only methods often suffer from reduced interpretability and generalization due to visual content variability.

The transition from unimodal to multimodal approaches has marked a significant evolution in automatic depression detection systems, especially in analyzing social media content. While unimodal systems process textual or visual data, multimodal models attempt to fuse both modalities, leveraging complementary emotional and contextual signals. For example, a social media post containing a cheerful image accompanied by a pessimistic caption may exhibit dissonance indicative of latent depressive symptoms. Architectures such as the Multimodal Feature Fusion Network (MFFNC) [18] and the Multimodal Object-Oriented Graph Attention Model (MOGAM) [44] have demonstrated that joint learning over text and images can significantly improve detection performance over unimodal baselines.

Despite these advancements, many current multimodal frameworks lack mechanisms for personalization. Depression manifests heterogeneously across individuals, influenced by their personality traits and behavioral dispositions. Recent research has highlighted the value of incorporating personality traits derived from the Myers–Briggs Type Indicator (MBTI) to enable user-specific modeling [9], [45]. For example, a recent study by Pradnyana et al. [9] demonstrated the positive contribution of MBTI-based personality information in enhancing the accuracy of depression detection from social media. The study employed an ensemble-based approach involving both average and meta-learning strategies, using RoBERTa and a hybrid Random Forest–BiLSTM architecture as the primary base learners. Several works have adopted ensemble learning strategies that combine predictions from modality-specific classifiers using methods like majority voting, averaging, or meta-learners [5], [19]. While ensemble methods offer interpretability and robustness, they typically assume that modalities operate independently and treat their fusion statically. The ensemble approach neglects modality interactions, where visual context can reframe textual meaning, and fails to adapt to dynamic shifts in modality relevance. These limitations suggest the need for deeper integration strategies. Cross-modal attention mechanisms offer an improvement by learning conditional relationships between modalities, enabling semantic alignment and intermodal reasoning.

### B. CROSS-MODAL ATTENTION MECHANISMS

Integrating attention mechanisms has significantly improved automatic depression detection, particularly in handling noisy and context-dependent data from social media [32]. Initially developed for natural language processing, attention mechanisms allow models to focus on the most informative parts of the input dynamically. In multimodal settings, cross-modal attention has emerged as a powerful strategy to model interactions between modalities such as text and images, enabling semantically aligned representations across different feature spaces [3], [4], [6], [18].

Cross-modal attention extends the standard self-attention mechanism by enabling one modality to be conditioned on another, such as using visual features to guide attention over text, or vice versa. This design is particularly relevant in social media contexts, where the semantic interpretation of a text caption is often shaped by its accompanying image. By aligning semantic cues across modalities, cross-modal attention enhances the model's ability to detect subtle indicators of depression that may be ambiguous or context-dependent when modalities are processed independently.

Li and Xiao [18] introduced the Multimodal Feature Fusion Network (MFFNC), which integrates MacBERT-derived textual embeddings with visual representations using a cross-attention transformer architecture. Their model achieved 94.95% accuracy on a benchmark dataset, outperforming conventional early and late fusion strategies. In a similar direction, Wang et al. [4] proposed CrossAMF, a unified multitask training framework based on cross-modal attention for multimodal depression detection. CrossAMF employs a sentiment-filtered XLNet-CNN-BiGRU pipeline for textual processing and a Reduce-VGGNet model for visual feature extraction, effectively fusing representations across modalities while preserving interpretability.

Further advances have explored hierarchical and multilevel fusion. For instance, Li et al. [3] designed a multistage cross-modal mechanism to integrate multimodal information iteratively, capturing the most salient joint features across heterogeneous data inputs. Iyortsuun et al. [6] developed the Additive Cross-Modal Attention Network (ACMA), which utilizes additive attention to compute inter-modal weights prior to fusion. While ACMA demonstrated potential in depression detection on benchmark datasets such as DAIC-WOZ and EATD-Corpus, it suffers from structural limitations that hinder its real-world applicability.

Specifically, ACMA implements a static fusion strategy, wherein modality-specific features (e.g., audio and text) are independently encoded and subsequently concatenated before passing through a shared additive attention module. This approach treats the combined multimodal input as a uniform vector, failing to capture dynamic interdependencies between modalities or adapt to varying modality relevance. Such design choices implicitly assume that each modality contributes equally across all instances, which is a limiting assumption, especially in social media environments where users exhibit high variability in their expressive behaviors. For example, some individuals may articulate depressive symptoms primarily through linguistic cues, while others may convey emotional distress more vividly through visual content. These limitations reinforce the need for adaptive and context-sensitive fusion mechanisms capable of dynamically weighting modalities based on their relative informativeness per user instance. The proposed direction toward adaptive gated fusion aims to address these challenges by introducing learnable gates that modulate modality contributions in a data-driven, instance-specific manner.

We propose DeXMAG, a personalized multimodal framework that integrates Cross-Modal Attention (CMA) and Adaptive Gated Fusion (AGF) to address these limitations. Unlike ACMA's static additive fusion, DeXMAG's CMA module explicitly models inter-modal dependencies, enabling mutual conditioning between textual and visual features to capture richer contextual signals. Furthermore, the AGF mechanism introduces dynamic weighting of modalities through a learned gating function, allowing the model to emphasize more informative modalities based on instance-specific cues adaptively. This fusion flexibility enhances the model's robustness to missing or noisy modalities and better aligns with the inherent variability of user-generated content. In addition, DeXMAG incorporates user-level personality trait attributes to enable personalized depression prediction. Together, these enhancements enable DeXMAG to overcome the fusion rigidity of ACMA and deliver improved performance in detecting depression from social media data.

## III. THE PROPOSED METHOD

This section provides a detailed explanation of the proposed DeXMAG model. As illustrated in Fig. 1, DeXMAG comprises a series of sequential stages designed to enable personalized depression detection from multimodal social media data. The model architecture consists of four primary components: (1) text feature extraction, (2) image feature extraction, (3) MBTI personality trait inference, and (4) modality fusion through a Cross-Modal Attention mechanism integrated with Adaptive-Gated Fusion. Each of these components is elaborated in Subsections A through C.

### A. DATA PREPROCESSING

Data preprocessing constitutes a fundamental stage in the DeXMAG model pipeline, serving to enhance the quality of multimodal social media data prior to feature extraction and model training. Given the heterogeneous nature of the input data, a comprehensive and modality-specific preprocessing protocol is essential.

#### 1) TEXTUAL DATA PREPROCESSING

In the proposed multimodal framework for personalized depression detection, textual information constitutes one of

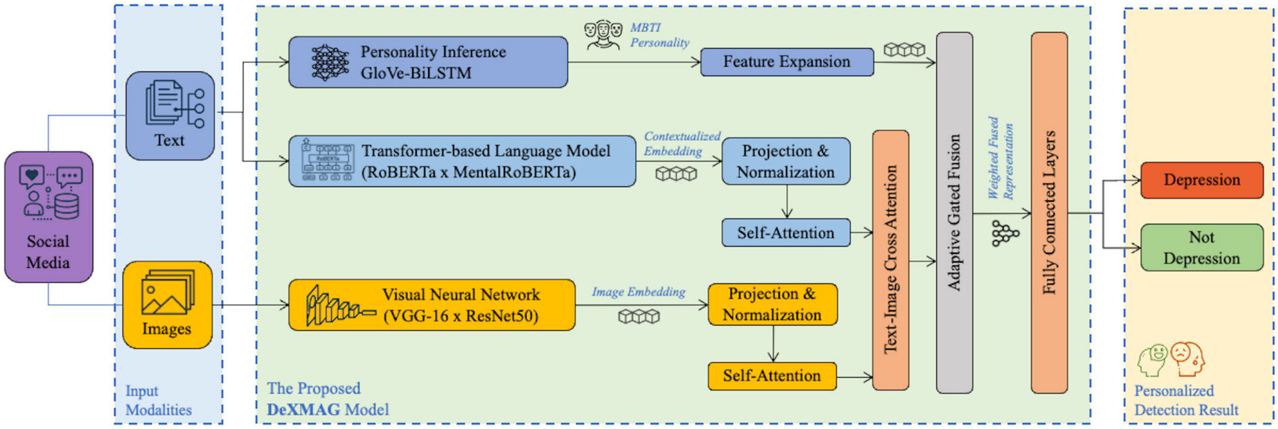

Social Media
Text
Images
Input Modalities
Personality Inference GloVe-BiLSTM
MBTI Personality
Feature Expansion
Transformer-based Language Model (RoBERTa x MentalRoBERTa)
Contextualized Embedding
Projection & Normalization
Self-Attention
Visual Neural Network (VGG-16 x ResNet50)
Image Embedding
Projection & Normalization
Self-Attention
Text-Image Cross Attention
Adaptive Gated Fusion
Weighted Fused Representation
Fully Connected Layers
The Proposed DeXMAG Model
Depression
Not Depression
Personalized Detection Result

**FIGURE 1.** Architecture of the proposed DeXMAG model for personalized depression detection from multimodal social media data.

the primary modalities. To address the informal, noisy, and heterogeneous nature of social media language, a comprehensive preprocessing pipeline was implemented to convert raw timeline data into a structured and normalized textual representation. The following preprocessing steps were applied:

- Timeline Text Extraction: Each user's timeline data was stored as a newline-delimited JSON stream, with each line representing an individual post. A custom extraction procedure was employed to parse each JSON entry, isolate the textual content, and concatenate posts into a continuous textual sequence per user. This aggregation preserves the narrative flow and behavioral context critical for personalized modeling.
- Lexical Normalization: Informal contractions commonly used in social media (e.g., "can't," "won't," "I'm") were expanded to their full lexical forms using a domain-specific contraction dictionary. This step reduces lexical sparsity and ensures consistency in token representation, improving the quality of contextual embeddings.
- Noise Reduction and Token Standardization: A series of regular expression rules were applied to anonymize sensitive content (e.g., replacing URLs with `<url>` and user mentions with `<user>`), normalize elongated words (e.g., "soooo" to "so"), and ensure syntactic clarity through punctuation separation. These transformations minimize irrelevant variance and enhance the stability of token distributions.
- Affective Emoji Mapping: Emoticons and Unicode-based emojis, which carry affective and psychological signals, were systematically replaced with interpretable textual tags (e.g., "<cry>" for crying face, "<smile>" for smiling face). This semantic enrichment ensures that emotional expressions embedded in visual symbols are preserved in the text modality.
- Text Standardization: Final cleanup operations addressed formatting inconsistencies such as excessive whitespace and residual non-standard characters. This step ensures a clean and uniform text input, optimized for tokenization and subsequent embedding generation.

This textual preprocessing pipeline is designed to yield semantically rich and emotionally grounded representations from noisy social media data, thereby enabling reliable feature extraction for personalized depression detection. The complementary visual modality is addressed in the subsequent section.

### 2) VISUAL DATA PREPROCESSING

Complementing the textual modality, visual data derived from user-uploaded images were also incorporated as part of the multimodal depression detection framework. These images, often expressive of affective states and lifestyle cues, were preprocessed to ensure compatibility with standard convolutional neural network (CNN) architectures and to minimize irrelevant variance in input representations. The following preprocessing steps were applied to each image:

- Image Loading and Conversion: Each image file was programmatically loaded and converted to the RGB color space to maintain consistency across inputs. This step ensures that all channels are standardized regardless of the original image format (e.g., grayscale, RGBA), thus preventing downstream inconsistencies during feature extraction.
- Resizing and Normalization: All images were resized to a fixed spatial resolution of 224 × 224 pixels, a standard input dimension for pre-trained CNNs such as VGG-16 and ResNet50. Following resizing, pixel values were normalized to the [0, 1] range by dividing by 255. This normalization facilitates stable model training and inference by reducing the dynamic range of input features.
- Error Handling: The preprocessing pipeline includes robust error handling to detect and skip over corrupted, invalid, or missing image files. This ensures data

integrity and prevents failures during batch processing or model training.

- Output Format: The final output consists of uniformly sized, three-channel, floating-point arrays representing the normalized visual content of each user image. These arrays were subsequently passed into deep convolutional backbones, either VGG-16 or ResNet50, for feature extraction.

This preprocessing approach ensures that raw visual data are transformed into a standardized format suitable for deep feature extraction, enabling the fusion of linguistic and visual information in downstream personalized depression detection tasks.

### B. TRANSFORMER-BASED LANGUAGE MODEL AS TEXTUAL FEATURE EXTRACTOR

In our proposed framework, text is utilized as the primary modality for depression detection, given its ubiquity and mandatory presence across most social media platforms. We adopt transformer-based language models (TLMs) pre-trained on large-scale corpora to extract rich and semantically meaningful textual representations from user-generated content. In this study, we compare the performance of a general-purpose model, the Robustly Optimized BERT Approach (RoBERTa) [46], with a domain-specific variant, MentalRoBERTa [13], to evaluate their respective capabilities in capturing depression-related linguistic cues within the DeXMAG framework.

RoBERTa, introduced by Liu et al. [46], is a transformer-based encoder that improves upon the original Bidirectional Encoder Representations from Transformers (BERT) [47] by eliminating the next sentence prediction objective, applying dynamic masking, and utilizing larger mini-batches and training sequences. Trained on general-domain corpora such as BookCorpus and Common Crawl, RoBERTa demonstrates strong performance in various natural language understanding tasks by learning robust semantic representations. Furthermore, we also incorporate MentalRoBERTa [13], a domain-adapted variant of RoBERTa that has undergone additional pretraining on mental health-related textual data sourced from online forums and platforms focused on psychological well-being. This continued pretraining enables MentalRoBERTa to more effectively capture domain-specific lexical features, emotional indicators, and cognitive patterns often associated with depressive symptoms and psychosocial distress. Fig. 2 illustrate the model architectures of MentalRoBERTa, highlighting its role in extracting deep textual features for integration into the multimodal DeXMAG model.

This study uses both RoBERTa and MentalRoBERTa exclusively as textual embedding extractors. Rather than employing the models for direct classification, we extract the contextualized representation of each input post via the final hidden state of the [CLS] token. This vector serves as a compact semantic embedding that encodes the overall meaning of the user's post. Fig. 2 illustrates the end-to-end process of extracting a semantic text representation from user-generated social media content using the MentalRoBERTa model, which serves as a domain-specific transformer-based language encoder in the DeXMAG architecture. The primary difference between RoBERTa and MentalRoBERTa lies in the pretraining data domain, while the underlying architecture remains the same.

The pipeline in Fig. 2 begins with a raw input sentence, referred to as a user-generated post. Let $x = w_1, w_2, \ldots, w_n$ be a user-generated post comprising a sequence of $n$ tokens. The MentalRoBERTa tokenizer maps each word $w_i$ into subword units using Byte-Pair Encoding (BPE), and prepends/appends special tokens $[CLS]$, $[SEP]$, and $[PAD]$ as needed. The resulting sequence is:

$$\widetilde{x} = [CLS], tok_1, tok_2, \ldots, tok_m, [SEP], \ldots, [PAD] \quad (1)$$

where $t_i \in \mathcal{V}$ where $\mathcal{V}$(the vocabulary), and $m \leq 512$ due to the model's maximum sequence length. An attention mask $A \in \{0, 1\}^{512}$ is created to distinguish valid tokens from padded positions, where:

$$A_i = \begin{cases} 1 & \text{if } \widetilde{x}_i \text{ is a valid token} \\ 0 & \text{if } \widetilde{x}_i = \texttt{[PAD]} \end{cases} \quad (2)$$

The tokenized input $\widetilde{x}$ is then passed through MentalRoBERTa's transformer encoder, which consists of 12 stacked self-attention layers. Each layer applies multi-head self-attention and feed-forward transformations to produce contextualized hidden representations:

$$H^{(l)} = \text{TransformerLayer}^{(l)}\left(H^{(l-1)}\right), \quad l = 1, \ldots, 12 \quad (3)$$

with the initial layer input $H^{(0)} \in \mathbb{R}^{512 \times d}$, where $d = 768$ is the hidden size. From the final encoder layer output $H^{(12)}$, the hidden state corresponding to the $[CLS]$ token is selected:

$$h_{[\text{CLS}]} = H_0^{12} \in \mathbb{R}^{768} \quad (4)$$

This vector $h_{[CLS]}$ serves as a compressed representation of the full input sequence and is referred to as the Textual Feature Vector. It is used as input to the DexMAG fusion block:

$$z_{text} = W \cdot h_{[cls]} + b \quad (5)$$

where $W \in \mathbb{R}^{d' \times 768}$ is a learnable projection matrix (for dimensional alignment) and $d'$ is the latent dimension for fusion with other modalities.

Including both models in our research enables a controlled investigation into the impact of domain-specific pretraining on depression detection performance. In this study, we implement the RoBERTa model using the PyTorch deep learning framework, integrated with the Transformers module provided by the HuggingFace library. After loading the pre-trained weights, we partially fine-tuned RoBERTa and MentalRoBERTa on our labelled depression dataset.

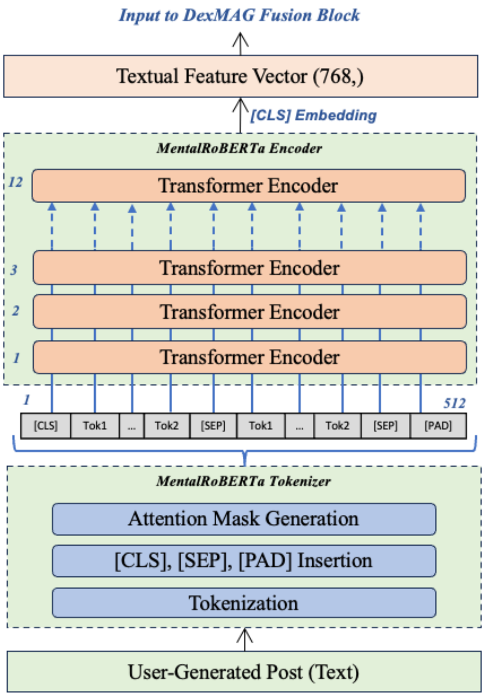


**FIGURE 2.** **MentalRoBERTa-based textual feature extraction pipeline for the DeXMAG model.**

### C. VISUAL NEURAL NETWORK AS VISUAL FEATURE EXTRACTOR

As illustrated in Fig. 1, visual data is incorporated as an additional modality to enhance depression detection from social media content. To extract semantically rich visual representations from user-shared images, we employ a Convolutional Neural Network (CNN)-based approach utilizing two widely recognized architectures: VGG-16 and ResNet-50. These models are selected due to their well-established performance in visual recognition tasks and their effectiveness in transfer learning scenarios. In this study, both models are used exclusively as feature extractors, with their classification layers removed, allowing us to integrate the extracted visual embeddings into the multimodal DeXMAG framework.

VGG-16 [48] and ResNet-50 [49] differ significantly in their architectural design and representational capacity. VGG-16 is a relatively deep network with a straightforward, sequential architecture composed of stacked convolutional layers with small receptive fields (3×3), followed by fully connected layers. It is known for capturing fine-grained spatial features but tends to be computationally intensive and memory-demanding due to the large number of parameters. In contrast, ResNet-50 introduces residual connections, allowing the network to train deeper layers without suffering from the vanishing gradient problem. Its bottleneck residual blocks enable more efficient gradient flow and parameter sharing, allowing it to learn more abstract and hierarchical features. Given these complementary characteristics, we compare the performance of VGG-16 and ResNet-50 to determine the most effective visual feature extractor for depression detection in the DeXMAG architecture. Fig. 4 provides a detailed illustration of the visual feature extraction process using (a) VGG-16 and (b) ResNet-50.

Let $I \in R^{H\times W\times 3}$ represent an RGB image associated with a user's social media post. To ensure compatibility with pretrained CNN architectures, each image is resized to a fixed dimension $I' \in R^{224\times 224\times 3}$. The image is then normalized according to the pre-processing scheme used in ImageNet-trained models. Once pre-processed, the image $I'$ is passed through either VGG-16 or ResNet-50, both pretrained on the ImageNet dataset. These models output high-dimensional feature maps from their penultimate convolutional layers, as presented in Fig. 4.

The VGG-16 model consists of 16-layer network composed of stacked convolutional and pooling layers, ending with three fully connected layers. For feature extraction, we truncate the model before the classification head and extract the activations from the last convolutional block:

$$f_{VGG} = Flatten\left(Conv5\_3\left(I'\right)\right) \in R^{d_v} \tag{6}$$

where $d_v$ ranges from 4,096 (flattened features) to lower dimensions with pooled. Meanwhile, ResNet-50 consist of 50-layer deep residual network composed of bottleneck residual blocks. The global average pooling layer before the classification head outputs:

$$f_{ResNet} = GAP(ResBlock4(I)') \in R^{2048} \tag{7}$$

where $GAP$ denotes Global Average Pooling applied to the final convolutional feature map.

Both extracted feature vectors $f_{VGG}$ and $f_{ResNet}$ capture abstract spatial patterns, object presence, and texture characteristics that might be relevant to psychological state. To ensure compatibility with the fusion layer, the high-dimensional visual features are projected into a common latent space:

$$Z_{image} = W_{img} \cdot f_{VNN} + b_{img}, Z_{image}e \in R^{d'} \tag{8}$$

where virtual neural network feature ($f_{VNN}$) is either $f_{VGG}$ or $f_{ResNet}$, $W_{img} \in R^{d'\times d_v}$ is a learnable projection matrix aligning the visual features with other modalities for downstream fusion, and $b_{img}$ refers to the bias vector in a linear transformation layer. This visual embedding $Z_{image}$ is then input into the Cross-Modal Attention and Adaptive-Gated Fusion block of our DeXMAG, where it interacts with textual and personality-based features to enable context-aware depression detection.

### D. MBTI PERSONALITY INFERENCE

Inspired by the study conducted by [9], this research incorporates user-level personality features into the depression detection framework based on the Myers–Briggs Type Indicator (MBTI). The MBTI is a widely recognized psychological model that categorizes personality into 16 distinct

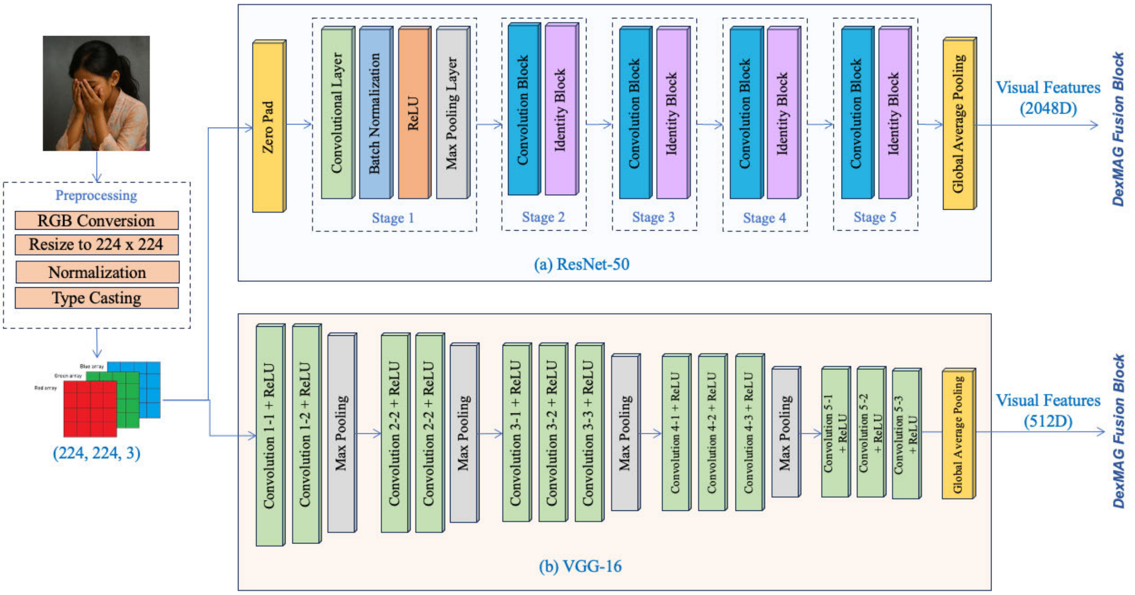


**FIGURE 3.** Visual feature extraction process in the DeXMAG model. (a) ResNet-50 and (b) VGG-16 are employed as VNN backbones to extract deep visual representations from user-shared images.

types derived from four dichotomous dimensions, as shown in Table 1. Each type represents a unique combination of cognitive preferences influencing behavior, communication style, and emotional expression.

**TABLE 1.** Overview of MBTI personality trait dimensions.

| Dimension | First Trait | Opposing Trait |
|---|---|---|
| **Energy Source** | **Extraversion (E)**: Gains energy from social interaction, active engagement, and external stimuli. | **Introversion (I)**: Gains energy from solitary activities, reflection, and internal focus. |
| **Information Processing** | **Sensing (S)**: Prefers concrete information, practical experience, and attention to detail. | **Intuition (N)**: Focuses on patterns, abstract concepts, and future possibilities. |
| **Decision Making** | **Thinking (T)**: Makes decisions using logical reasoning, objective analysis, and fairness. | **Feeling (F)**: Makes decisions based on values, empathy, and consideration for others. |
| **Lifestyle Orientation** | **Judging (J)**: Prefers structure, planning, and organization. Values closure and clear outcomes. | **Perceiving (P)**: Prefers flexibility, spontaneity, and adaptability. Comfortable with uncertainty. |

We employed a cross-dataset personality inference strategy to obtain MBTI personality information for users in the depression dataset. This process involves adapting a personality classification model trained on an MBTI-labelled dataset to a different target domain with similar data characteristics, namely the depression dataset. In this study, we used the MBTI dataset from Kaggle to demonstrate the feasibility of personality inference. The dataset was collected from the Personality Cafe forum, a public online community focused on personality self-discovery and typological discussion. To the best of the authors' knowledge, there is no unique standard dataset for machine-learning techniques based on the MBTI instrument. The dataset specifications are described in detail in Section IV.

The adapted model used in this study is a GloVe-BiLSTM network, originally proposed and validated in prior research [50]. The model was trained to learn semantic and syntactic patterns in user-generated content using GloVe word embeddings and a bidirectional long short-term memory architecture, effectively capturing linguistic cues associated with MBTI personality traits. Once trained, the GloVe-BiLSTM model was applied to the depression dataset to predict personality profiles for each user. Instead of assigning hard MBTI labels, the model produces probabilistic scores for each trait pair (I/E, N/S, T/F, J/P), where the output values range from 0 to 1. These scores indicate the predicted likelihood of a user exhibiting a particular personality trait along each dimension. For example, a score closer to 0 in the I/E dimension suggests a higher probability of extraversion, while a score closer to 1 indicates a stronger tendency toward introversion. This probabilistic representation provides a continuous and fine-grained personality characterization, allowing the depression detection model to capture nuanced user-specific traits rather than relying on rigid categorical labels.

### E. CROSS-MODAL ATTENTION WITH ADAPTIVE GATED FUSION

To effectively integrate heterogeneous modalities for personalized depression detection, we propose a novel fusion mechanism that combines Cross-Modal Attention (CMA) with Adaptive Gated Fusion (AGF), enhanced by personality-aware conditioning. The fusion block receives three input streams: textual features, visual features, and user personality profiles encoded as four-dimensional MBTI-based probabilities. The architecture is optimized for personalized depression detection by enabling dynamic modulation of modality relevance for each user. The detailed steps of the fusion process are presented in Algorithm 1.

Textual features extracted via RoBERTa or Mental-RoBERTa and image features extracted from either VGG-16 or ResNet50 are first projected into a shared 512-dimensional latent space using dense layers with L2 regularization. Given textual features $T \in R^{d_t}$, visual features $V \in R^{d_v}$, and MBTI-based personality traits $P \in R^4$, we first project all inputs into a shared latent space of 512 dimensions using (9) and (10).

$$T_{proj} = \text{LayerNorm}(\text{ReLU}(W_t T + b_t)) \tag{9}$$

$$V_{proj} = \text{LayerNorm}(\text{ReLU}(W_v V + b_v)) \tag{10}$$

Each modality is then subjected to intra-modal self-attention. The attention weights $\alpha_T, \alpha_V \in \mathbb{R}^{512}$ are computed using a dense transformation with a tanh activation, followed by a softmax normalization:

$$\alpha_T = \text{Softmax}(\tanh(W_{\alpha T} T_{proj} + b_{\alpha T})) \tag{11}$$

$$\alpha_V = \text{Softmax}(\tanh(W_{\alpha V} V_{proj} + b_{\alpha V})) \tag{12}$$

$$T_{att} = T_{proj} \odot \alpha_T \tag{13}$$

$$V_{att} = V_{proj} \odot \alpha_V \tag{14}$$

To capture semantic alignment between modalities, we compute cross-modal attention by allowing each modality to attend to the other's self-attended representation:

$$\beta_T = \text{Softmax}(\tanh(W_{\beta T} V_{att} + b_{\beta T})) \tag{15}$$

$$\beta_V = \text{Softmax}(\tanh(W_{\beta V} T_{att} + b_{\beta V})) \tag{16}$$

$$T_{cross} = T_{att} \odot \beta_T \tag{17}$$

$$V_{cross} = V_{att} \odot \beta_V \tag{18}$$

The personality vector $P$ is expanded via a multi-layer nonlinear transformation (19) to match the 512-dimensional latent space, ensuring semantic compatibility.

$$P_{exp} = \tanh(W_p^3(\tanh(W_p^2(\tanh(W_p^1 P + b_p^1)) + b_p^2)) + b_p^3) \tag{19}$$

To allow the model to dynamically modulate the contribution of each $T, V, P$ modality, we introduce adaptive gates $g_T$, $g_V$, and $g_P$ computed through independent sigmoid-activated dense layers:

$$g_T = \sigma(W_{gT} T_{cross} + b_{gT}) \tag{20}$$

$$g_V = \sigma(W_{gV} V_{cross} + b_{gV}) \tag{21}$$

$$g_P = \sigma(W_{gP} P_{exp} + b_{gP}) \tag{22}$$

As in (23), (24), and (25), each modality is scaled by its corresponding gate, resulting in weighted modality representations.

$$T_{weighted} = g_T \odot T_{cross} \tag{23}$$

$$V_{weighted} = g_V \odot V_{cross} \tag{24}$$

$$P_{weighted} = g_P \odot P_{exp} \tag{25}$$

These representations are fused through summation, batch-normalized, and regularized via dropout:

$$F_{fused} = \text{Dropout}(\text{BatchNorm}(T_{weighted} + V_{weighted} + P_{weighted})) \tag{26}$$

Finally, the fused vector is passed through two dense layers for high-level abstraction before being classified with a sigmoid-activated output unit:

$$F_1 = \text{Dropout}(\text{ReLU}(W_1 F_{fused}) + b_1) \tag{27}$$

$$F_2 = \text{Dropout}(\text{ReLU}(W_2 F_{fused}) + b_2) \tag{28}$$

$$\hat{y} = \sigma(W_o F_2 + b_o) \tag{29}$$

This architecture facilitates flexible and user-aware multimodal reasoning, thereby enhancing the model's capacity to identify depressive signals across a wide spectrum of expressive behaviors exhibited by users on social media platforms. The detailed sequence of the proposed fusion process within the DeXMAG model is outlined in Algorithm 1.

## IV. EXPERIMENTAL EVALUATION

This section presents a comprehensive analysis of our proposed DexMAG model through a series of experiments. We begin by detailing the datasets and experimental setup, including the data specifications, training configurations, and evaluation metrics used to assess model performance. The second part reports empirical results, highlighting the relationship between MBTI personality traits and depression. Subsequently, we conduct ablation studies to assess the contribution of each input modality and personality information to overall performance. The effectiveness of the proposed model is benchmarked against several state-of-the-art (SOTA) baselines using rigorous statistical tests. Finally, a visualization of detection results is provided to illustrate the model's ability to distinguish depressive from non-depressive users based on multimodal and personalized cues.

### A. DATASETS AND EXPERIMENT SETUP

This study employs two distinct datasets derived from social media platforms to facilitate multimodal and personalized depression detection. The primary dataset is the depression-labeled dataset. The secondary dataset is the MBTI model personality dataset used during the cross-dataset personality inference. Although originally independent, both datasets share similar structural and linguistic characteristics because they originate from social media content.

The primary dataset used in this research is the multimodal depression dataset introduced by Gui et al. [21]. It comprises user-generated content collected from the Twitter platform and includes 6,562 users. These users are categorized into two groups. There are 1,402 users identified as depressed and 5,160 as non-depressed. Depression labels were assigned based on user self-disclosures. Users in the depressed group explicitly referenced clinical diagnoses or related terminology in their posts. In contrast, users in the

**Algorithm 1** Cross-Modal Attention With Adaptive Gated Fusion (CMA-AGF)

```
Require: Text features $\mathbf{T}$, image features $\mathbf{V}$, personality vector $\mathbf{P}$
Ensure: Depression prediction score $\hat{y}$
// Projection and Normalization
$\mathbf{T}_{\text{proj}} \leftarrow \text{LayerNorm}(\text{Dense\_ReLU}(\mathbf{T}, \text{units} = 512))$
$\mathbf{V}_{\text{proj}} \leftarrow \text{LayerNorm}(\text{Dense\_ReLU}(\mathbf{V}, \text{units} = 512))$
// Self-Attention within Modalities
$\boldsymbol{\alpha}_T \leftarrow \text{Softmax}(\text{Dense\_Tanh}(\mathbf{T}_{\text{proj}}))$
$\boldsymbol{\alpha}_V \leftarrow \text{Softmax}(\text{Dense\_Tanh}(\mathbf{V}_{\text{proj}}))$
$\mathbf{T}_{\text{att}} \leftarrow \mathbf{T}_{\text{proj}} \odot \boldsymbol{\alpha}_T$
$\mathbf{V}_{\text{att}} \leftarrow \mathbf{V}_{\text{proj}} \odot \boldsymbol{\alpha}_V$
// Cross-Modal Attention
$\boldsymbol{\beta}_T \leftarrow \text{Softmax}(\text{Dense\_Tanh}(\mathbf{V}_{\text{att}}))$
$\boldsymbol{\beta}_V \leftarrow \text{Softmax}(\text{Dense\_Tanh}(\mathbf{T}_{\text{att}}))$
$\mathbf{T}_{\text{cross}} \leftarrow \mathbf{T}_{\text{att}} \odot \boldsymbol{\beta}_T$
$\mathbf{V}_{\text{cross}} \leftarrow \mathbf{V}_{\text{att}} \odot \boldsymbol{\beta}_V$
// Personality Expansion
$\mathbf{P}_{\text{exp}} \leftarrow \text{Dense\_Tanh}(\mathbf{P}, 128)$
$\mathbf{P}_{\text{exp}} \leftarrow \text{Dense\_Tanh}(\mathbf{P}_{\text{exp}}, 256)$
$\mathbf{P}_{\text{exp}} \leftarrow \text{Dense\_Tanh}(\mathbf{P}_{\text{exp}}, 512)$
// Adaptive Gating
$g_T \leftarrow \text{Sigmoid}(\text{Dense}(\mathbf{T}_{\text{cross}}))$
$g_V \leftarrow \text{Sigmoid}(\text{Dense}(\mathbf{V}_{\text{cross}}))$
$g_P \leftarrow \text{Sigmoid}(\text{Dense}(\mathbf{P}_{\text{exp}}))$
$\mathbf{T}_{\text{weighted}} \leftarrow g_T \odot \mathbf{T}_{\text{cross}}$
$\mathbf{V}_{\text{weighted}} \leftarrow g_V \odot \mathbf{V}_{\text{cross}}$
$\mathbf{P}_{\text{weighted}} \leftarrow g_P \odot \mathbf{P}_{\text{exp}}$
// Fusion
$\mathbf{F}_{\text{fused}} \leftarrow \mathbf{T}_{\text{weighted}} + \mathbf{V}_{\text{weighted}} + \mathbf{P}_{\text{weighted}}$
$\mathbf{F}_{\text{drop}} \leftarrow \text{Dropout}(\text{BatchNorm}(\mathbf{F}_{\text{fused}}), \text{rate} = 0.6)$
// Final Dense Layers and Prediction
$\mathbf{F}_{\text{dense}} \leftarrow \text{Dropout}(\text{Dense\_ReLU}(\mathbf{F}_{\text{drop}}, \text{units} = 512, \text{L2}))$
$\mathbf{F}_{\text{dense}} \leftarrow \text{Dropout}(\text{Dense\_ReLU}(\mathbf{F}_{\text{dense}}, \text{units} = 256))$
$\hat{y} \leftarrow \text{Sigmoid}(\text{Dense}(\mathbf{F}_{\text{dense}}, \text{units} = 1))$
return $\hat{y}$
```

non-depressed group did not express any indications of depressive symptoms. This labeling scheme is formulated as a binary classification task, distinguishing between depressed and non-depressed users. However, it is important to emphasize that the depression labels used in this study should be interpreted as proxies for depressive expression patterns rather than definitive diagnostic outcomes. Users may self-identify with depressive experiences without a formal clinical assessment, and conversely, individuals with diagnosed depression may choose not to disclose their condition due to stigma or privacy concerns. As such, the dataset reflects linguistic and behavioral signals associated with depression in naturalistic social media contexts, rather than clinically validated diagnoses.

Each user profile contains textual content and visual data. In cases where image data is unavailable, a zero vector is used to represent the visual modality. This process ensures consistency in the multimodal input structure across all samples. The depression dataset is the main resource for training and evaluating the proposed depression detection model. User-level personality features were integrated through a cross-dataset inference process to incorporate personalized characteristics.

This study employs the Myers–Briggs Type Indicator (MBTI) Personality Dataset to obtain personality representations. This dataset is widely used in social media-based personality modeling, as shown by recent research conducted by [50], [51], and [52]. It contains 8,675 instances, each comprising fifty sentences drawn from individual social media posts and a corresponding MBTI label. The MBTI types follow the standard 16-type classification based on four psychological dimensions. The label distribution is as follows. There are 6,676 instances labeled as Introversion and 1,999 as Extraversion. For the second dimension, there are 1,197 labeled as Intuition and 7,478 as Sensing. The third dimension includes 4,694 labeled as Thinking and 3,981 as Feeling. The final dimension includes 5,241 labeled as Judging and 3,434 as Perceiving. This dataset trains a personality classification model that learns to map user-generated content to MBTI types. This process, referred to as cross-dataset personality inference, enables the enrichment of the depression dataset with personality-informed features.

**TABLE 2. Summary statistics of the dataset employed in our experiments. The columns #*Text* and #*Text* + *Image* indicate the number of tweets containing only textual content and those comprising paired text and image data, respectively.**

| Category | No. of Users | #Text | #Text+Image |
|---|---:|---:|---:|
| Depressed | 1,402 | 251,834 | 40,730 |
| Non-Depressed | 5,160 | 3,302,366 | 650,817 |

To assess the model's effectiveness in detecting depression, we employed standard classification metrics: accuracy (Acc), precision, recall, and F1-score. Together, these metrics provide a nuanced understanding of the model's diagnostic capability for real-world depression detection scenarios. Hyperparameter tuning was conducted via grid search on the validation set to optimize model performance. We focused on three primary parameters: epoch, batch size, and learning rate. For our experiments, we tested epochs in the range of 100–200, batch sizes of 32 and 64, and a fixed learning rate of 1e-4, as summarized in the parameter column of Table 3. Eight experimental scenarios were conducted to compare different base combinations of textual and visual encoders, with and without the Adaptive Gated Fusion (AGF) mechanism. For text encoding, we compared RoBERTa-base and MentalRoBERTa-base, while for visual encoding, we evaluated both VGG-16 and ResNet-50 architectures. The initial four scenarios involved basic multimodal fusion

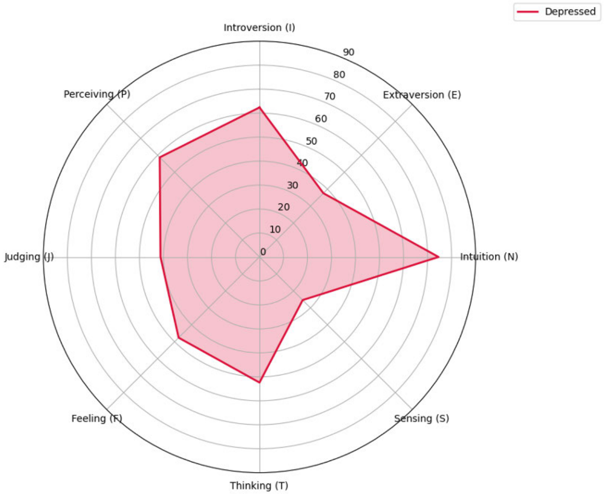


**FIGURE 4.** Radar chart illustrating the distribution of MBTI personality traits among users identified as depressed. The values represent the percentage of individuals exhibiting each trait within the depressed group.

without AGF, serving as baselines. The remaining four scenarios incorporated AGF to enable dynamic modality weighting. Furthermore, a series of ablation studies were performed to systematically examine the contribution of each input modality (text, image) and auxiliary features (MBTI personality traits).

All experiments in this study were conducted using a high-performance computing environment equipped with an NVIDIA A100 Tensor Core GPU. We utilized the Hugging Face Transformers library, which provides pre-trained models and tools for natural language processing tasks, including the RoBERTa and MentalRoBERTa variants used in this study. For visual feature extraction, we employed the VGG and ResNet architectures, both implemented using the Keras and TensorFlow deep learning frameworks. These libraries facilitated the integration of textual and visual modalities, enabling end-to-end training of the proposed multimodal depression detection model

## B. RESULTS AND DISCUSSION

### 1) RELATIONSHIP BETWEEN MBTI PERSONALITY TRAITS AND DEPRESSION IN SOCIAL MEDIA USERS

In this section, we further investigate the relationship between users' personality characteristics, as inferred through MBTI personality trait extraction, and the presence of depressive tendencies within the dataset. The analysis of MBTI personality traits and their relationship to depressive tendencies is grounded in the theoretical structure of the Myers–Briggs Type Indicator (MBTI), which classifies individuals into 16 personality types based on unique combinations of four dichotomous trait pairs: Introversion–Extraversion (I/E), Sensing–Intuition (S/N), Thinking–Feeling (T/F), and Judging–Perceiving (J/P). Each personality type represents a distinct cognitive and behavioral style derived from the interaction of these four preference dimensions. To capture both global and specific personality patterns among depressed users, we employed two complementary visualizations. To provide a broader perspective, the radar chart shown in Fig. 4 illustrates the distribution of MBTI personality traits among users labeled as depressed. The contribution percentage denotes the proportion of depressed individuals exhibiting a particular trait relative to the total number of depressed users, thereby reflecting the prevalence of specific personality patterns associated with depression.

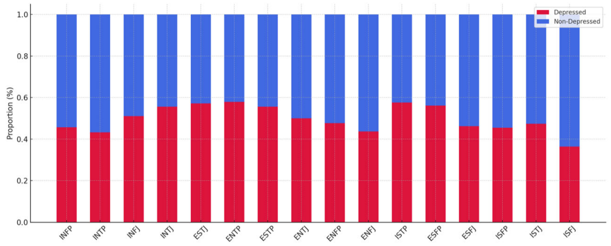


**FIGURE 5.** Proportional distribution of depressed and non-depressed users within each MBTI personality type. Each bar represents 100% of users for a given type, partitioned by depression label.

The distribution of MBTI personality traits among depressed users, as shown in Fig. 4, reveals significant patterns. Introversion (I) and Intuition (N) traits exhibited the highest contribution rates, with approximately 63% and 82% of depressed users displaying these characteristics, respectively. This contribution suggests that individuals who are inwardly focused and attuned to abstract possibilities may be more vulnerable to depressive symptoms. In contrast, Extraversion (E) and Sensing (S) traits are markedly underrepresented in the depressed cohort. These traits are associated with external engagement, present-oriented cognition, and practical problem-solving, which may serve as psychological buffers against emotional dysregulation. Furthermore, a higher prevalence of Perceiving (P) traits (59%) relative to Judging (J) traits (40%) suggests that individuals who favor spontaneity and adaptability over structured routines may be more vulnerable to stress accumulation and decision fatigue, potentially heightening their risk of depression.

These findings are consistent with prior psychological literature, which has reported strong associations between personality dimensions and vulnerability to mental health disorders, particularly depression [34], [35], [37]. Psychologically, introverted individuals may be more prone to internalizing affective disturbances. In contrast, due to their emphasis on hypothetical or future-oriented thinking, intuitive individuals may experience increased cognitive load related to uncertainty and existential concerns. Feeling (F) types prioritize interpersonal harmony and emotional resonance, which may render them more susceptible to emotional turbulence, whereas Thinking (T) types often adopt a more analytical approach to adversity. Likewise, individuals with a Perceiving (P) preference may exhibit reduced behavioral regulation and lower stress tolerance in unstructured environments, unlike Judging (J) individuals who rely on organization and planning to manage emotional challenges.

However, trait-level analysis alone cannot fully capture the nuanced interplay among dimensions that define MBTI personality types. Therefore, to extend our analysis, Fig. 5 presents a normalized bar chart showing the proportion of depressed and non-depressed individuals within each of the 16 MBTI types. The normalized distribution presented in Fig. 5 offers an information of depression prevalence across MBTI personality types by displaying the proportion of depressed users within each type, thus controlling for base frequency. Notably, types such as INTJ, ENTP, and ESTJ exhibit the highest internal proportions of depressed users, each exceeding 55%. This finding suggests that users with these personality types may possess latent psychological configurations that increase susceptibility to depressive tendencies despite not being the most frequent in the overall dataset.

These results in Fig. 5 complement the trait-level analysis depicted in Fig. 4, where Introversion (I), Intuition (N), and Perceiving (P) emerged as dominant traits among depressed users. INTJ and ENTP include the Intuition (N) component, while INTJ aligns with Introversion (I). ESTJ incorporates Judging (J), suggesting that while intuitive and introverted traits are indicative of depression risk, high proportions of depression can also manifest in structured (J) or extroverted (E) configurations, implying that other latent cognitive-emotional mechanisms may contribute beyond simple trait prevalence. This nuanced observation aligns with psychological theories emphasizing that depression is multifactorial, influenced not just by dominant traits but also by how these traits interact with cognitive coping styles, stress processing, and emotional regulation strategies.

Interestingly, MBTI types such as INFP and INTP, which appeared as the most frequent among depressed users in the raw distribution, show more balanced depressed-to-non-depressed ratios in this normalized chart. This indicates that although individuals with these types are numerous in the dataset, their relative susceptibility to depression may be average rather than elevated. This insight reinforces the importance of normalization when evaluating personality-depression relationships, preventing misinterpretation due to skewed population baselines.

Conversely, ISFJ and ISTJ demonstrate lower depression proportions, suggesting that their dominant traits, Sensing (S) and Judging (J), may offer protective mechanisms through grounded, present-focused perception and preference for structure and routine. This observation aligns with existing psychological studies [34], [35] that identify structured decision-making (J) and concrete, sensory awareness (S) as factors associated with lower emotional reactivity and greater resilience under psychological stress.

Together, the trait-level (Fig. 4) and type-level (Fig. 5) analyses offer a holistic understanding of how personality contributes to depression vulnerability. They illustrate that individual traits and interactions in complete type configurations are crucial for identifying at-risk individuals. These insights strengthen the rationale for integrating MBTI-based personality features into our proposed personalized multimodal depression detection model, allowing for more adaptive, psychologically informed, and user-specific predictions.

### 2) PERFORMANCE EVALUATION OF THE PROPOSED DeXMAG MODEL

We evaluate the effectiveness of our proposed personalized multimodal depression detection framework, DeXMAG, by benchmarking its performance against multiple baseline and state-of-the-art (SOTA) models. Through a series of systematically designed experiments and ablation studies, we demonstrate that DeXMAG consistently outperforms existing methods regarding classification accuracy, precision, recall, and F1-score. The results validate the model's ability to leverage cross-modal cues and user-specific personality traits to enhance the robustness and personalization of depression detection in social media contexts.

In the first set of experiments, we evaluated the performance of the proposed DeXMAG model across different combinations of modality-specific feature extractors. We assessed the effectiveness of the Adaptive Gated Fusion (AGF) mechanism. Table 3 shows that the baseline models combine either RoBERTa-base or MentalRoBERTa-base for textual feature extraction with either VGG-16 or ResNet-50 for visual modality encoding. The results indicate that MentalRoBERTa consistently outperforms RoBERTa across all visual backbone configurations, demonstrating the advantage of using domain-specific language models pretrained on mental health-related corpora. For example, the MentalRoBERTa + VGG-16 configuration achieved an F1-score of 0.9471, surpassing the RoBERTa + VGG-16 configuration (F1 = 0.9301) under identical settings.

Introducing the AGF mechanism further improves performance across all base combinations by enabling dynamic weighting of text and image modalities based on contextual relevance. This improvement is particularly evident in the MentalRoBERTa + VGG-16 + AGF configuration, which yields the highest overall performance with an accuracy of 0.9520, precision of 0.9617, recall of 0.9571, and F1-score of 0.9594. These results suggest that the AGF module successfully learns to modulate modality contributions in a context-aware manner, thereby enhancing robustness across diverse user-generated content. Additionally, the combination of VGG-16 and MentalRoBERTa synergizes effectively with AGF, outperforming its ResNet-50 counterpart despite ResNet's deeper architecture. This result may be attributed to VGG-16's simpler and more regular feature representation, which aligns better with the input content's relatively abstract and emotional nature.

Overall, these findings confirm the critical role of both modality configuration and the adaptive fusion strategy in enhancing the performance of multimodal depression detection systems. The superior results obtained by DeXMAG with

**TABLE 3.** Performance comparison of different combinations of modality-specific feature extractors used in the DeXMAG framework. The table reports Accuracy, Precision, Recall, and F1-Score across configurations combining RoBERTa-base or MentalRoBERTa-base with either VGG-16 or ResNet50, with and without Adaptive Gated Fusion (AGF).

| Feature Extractor Combination | Accuracy | Precision | Recall | F1-Score | Parameters |
|---|---|---|---|---|---|
| RoBERTa-base and VGG-16 | 0.9181 | 0.9415 | 0.9190 | 0.9301 | Ep=100; bz=32, lr=1e-4 |
| RoBERTa-base and ResNet50 | 0.9040 | 0.9444 | 0.8905 | 0.9167 | Ep=100; bz=32, lr=1e-4 |
| MentalRoBERTa-base and VGG-16 | 0.9379 | 0.9563 | 0.9381 | 0.9471 | Ep=200; bz=64, lr=1e-4 |
| MentalRoBERTa-base and ResNet50 | 0.9294 | 0.9502 | 0.9190 | 0.9392 | Ep=200; bz=64, lr=1e-4 |
| RoBERTa-base and VGG-16 w/ AGF | 0.9237 | 0.9378 | 0.9333 | 0.9356 | Ep=100; bz=32, lr=1e-4 |
| RoBERTa-base and ResNet50 w/ AGF | 0.9209 | 0.9213 | 0.9476 | 0.9343 | Ep=100; bz=32, lr=1e-4 |
| MentalRoBERTa-base and VGG-16 w/ AGF | 0.9520 | 0.9617 | 0.9571 | 0.9594 | Ep=200; bz=64, lr=1e-4 |
| MentalRoBERTa-base and ResNet50 w/ AGF | 0.9407 | 0.9565 | 0.9429 | 0.9496 | Ep=200; bz=64, lr=1e-4 |

**TABLE 4.** Performance comparison of depression detection approaches utilizing either single modality (text only) or multimodality (text and image), evaluated across four metrics: accuracy, precision, recall, and F1-score. The proposed DeXMAG model, which integrates textual, visual, and personality-aware features, demonstrates the highest performance across all metrics, highlighting the benefit of incorporating both multimodal and personalized information.

| **Model** | **Modality** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|---|---|
| CNN-BiLSTM Attention [40] | Text | 0.922 | 0.928 | 0.924 | 0.926 |
| MentalRoBERTa [13] | Text | 0.915 | 0.918 | 0.916 | 0.917 |
| DepRoBERTa [42] | Text | 0.903 | 0.905 | 0.905 | 0.905 |
| Avg-RoBERTforBi [9] | Text (+personality feature) | 0.928 | 0.928 | 0.928 | 0.928 |
| MTAL [53] | Text and Image | 0.896 | 0.897 | 0.897 | 0.897 |
| GRU + VGG-Net + COMMA [21] | Text and Image | 0.902 | 0.904 | 0.900 | 0.902 |
| Time2VecTransformer [28] | Text and Image | 0.934 | 0.936 | 0.932 | 0.934 |
| Attention + AlBERT-VGG16 [32] | Text and Image | 0.927 | 0.927 | 0.927 | 0.927 |
| The Proposed DeXMAG Model | Text and Image (+personality feature) | **0.952** | **0.961** | **0.957** | **0.959** |

the AGF mechanism underscore its capacity to dynamically leverage complementary information from heterogeneous modalities, effectively modulating the relevance of text and image inputs based on contextual cues. Moreover, the integration of personality-aware modeling further contributes to personalization, improving the model's ability to adapt to individual differences in expressive behavior.

Furthermore, we conducted a comparative evaluation to benchmark the performance of our proposed DeXMAG model against existing state-of-the-art (SOTA) approaches for depression detection on social media. These baseline models include unimodal frameworks, which utilize textual features independently, and multimodal frameworks combining text and image inputs. The comparison aims to assess the ability of DeXMAG to outperform prior methods by effectively integrating cross-modal information and incorporating user-specific personality traits. Performance is measured using accuracy, precision, recall, and F1-score to ensure a comprehensive evaluation across detection sensitivity and reliability.

Table 4 presents a comparative analysis of various depression detection models utilizing either single-modality (text) or multimodality (text and image), with some incorporating personality-aware features. The performance is evaluated across four metrics: accuracy, precision, recall (calculated from precision and F1-score), and F1-score.

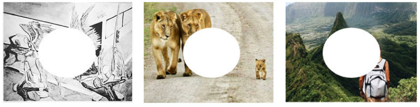
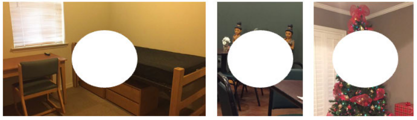
Among the text-only models, the CNN-BiLSTM Attention model [40] achieved an accuracy of 0.922 and an F1-score of 0.926, outperforming both MentalRoBERTa [13] and DepRoBERTa [42], which yielded slightly lower scores. The integration of personality features, as seen in Avg-RoBERTforBi [9], improved performance over baseline textual models, attaining an accuracy and F1-score of 0.928, suggesting that personality-aware modeling contributes positively to personalized depression detection.

In contrast, multimodal approaches that incorporate both text and image modalities exhibit generally higher performance than their unimodal counterparts. Notably, Time2VecTransformer [28] and Attention + AlBERT-VGG16 [32] reported strong F1-scores of 0.934 and 0.927, respectively. However, the proposed DeXMAG model, which further integrates personality-aware features alongside text and image, significantly outperformed all other models across all metrics, achieving an accuracy of 0.952, precision of 0.961, recall of 0.957, and an F1-score of 0.959.

These findings underscore the effectiveness of incorporating both multimodal content and user-specific personality signals for enhancing depression detection. The superior performance of DeXMAG indicates that the adaptive integration of heterogeneous features enables the model to better capture nuanced and latent indicators of depression that might be overlooked in unimodal or non-personalized approaches.

While DeXMAG was developed and evaluated using English-language social media data, the proposed architecture is inherently modular and not dependent on any specific language. In particular, the core components, such as the cross-modal fusion strategy and the visual and personality-aware modules, can theoretically operate across different languages because they are designed to be language agnostic. This means the model can be adapted to other languages,

**TABLE 5.** Representative examples illustrating the comparative performance of the proposed DeXMAG model and the previous model, including textual, visual, and personality-based inputs.

| Username | Content | True Label | Previous Model | The Proposed DeXMAG Model | |
|---|---|---|---|---|---|
| | | | | **MBTI Trait Probability** | **Detection Result** |
| @*Vis**y | **Textual:**<br>... Flip side of coin the mirror part of me just attaches to any type of energy I surround myself with. This has been very dangerous in past. When I'm around great crowd or myself I Am determined to be best I can be mind body spirit... I Am a perfectionist and a mirror at the same time... This why I must be by myself or around a great crew. Listen to COUNTRY BUS RIDDIM – Great mashup lots of good words <smile> <url> you care for the Earth you care for the Heart... Proverbs 10:30 The Righteous shall never be removed: but the wicked shall not inhibit the Earth. Woke up smiling, stood in shower dancing... Ready to take on today with the power of Light and Love...<br>**Visual:**<br><br>*(Three representative images sampled from a total of 62 user-uploaded images)* | Depression | Not Depression | IE: 0.510 (Introversion)<br>NS: 0.545 (Intuition)<br>TF: 0.529 (Thinking)<br>JP: 0.283 (Perceiving) | Depression |
| @Ha***_Rh***** | **Textual:**<br>... I hate myself... <user> "I'd prefer bangs to a seven head." "Excuse you, all of the books are neatly shelved, and all the patrons are happy." <url> my friends just seem so happy and I'm so happy for them and I just <heart>. RT <user> The only math I need in my life #WelcomeBackEd <url> "you shut your mouth." RT <user> "Beck and Jade will forever be my favorite tv show couple" <url> <user> "my phone did that and I am feeling personally attacked rn." <user> "yo my phone hates me." "I'm counting NYE <joy>" <url> "should not have even been tagged in it lol mal" <url> <user> "I is not sayin' you all fake, but I went out of town twice this week and you all got together both times"...<br>**Visual:**<br><br>*(Three representative images sampled from a total of 7 user-uploaded images)* | Not Depression | Depression | IE: 0.128 (Extroversion)<br>NS: 0.793 (Intuition)<br>TF: 0.771 (Thinking)<br>JP: 0.546 (Judging) | Not Depression |

provided that the textual encoder is replaced or fine-tuned using language-appropriate resources. To extend DeXMAG's applicability to low-resource language settings, future work may incorporate multilingual pretrained models such as XLM-RoBERTa or mBERT, apply cross-lingual transfer learning, or use culturally adapted training data. Moreover, the inclusion of visual and personality-based features offers additional signals that may partially mitigate limitations posed by scarce or low-quality textual data. These directions are essential for supporting equitable and culturally inclusive applications of mental health detection technologies.

### 3) DETECTION ILLUSTRATION

Table 5 presents a representative example that highlights the improved detection capability of the proposed model compared to existing state-of-the-art methods. To ensure user privacy when presenting qualitative examples from the depression detection dataset, we applied several privacy-preserving techniques. Usernames were partially masked using asterisks (*), and only selected segments of textual posts were shown with ellipses (...) to obscure the full content. For visual data, three randomly cropped images were included per user.

In the illustrative example in Table 5, the proposed DeXMAG model demonstrates superior performance by accurately identifying depressive tendencies subtly embedded within spiritually framed and motivational language. In the first sample, the previous model misclassified the instance as non-depressed, likely due to its reliance on superficial sentiment analysis and the presence of seemingly positive expressions such as "*the power of Light and Love*" and "*mind body spirit.*" While such spiritual language commonly reflects themes of personal growth, resilience, and metaphysical insight, it can also obscure latent psychological distress, thereby complicating depression detection through conventional models. DeXMAG effectively addresses this limitation by integrating multimodal signals to uncover latent indicators of depression.

The user's MBTI-based personality traits further contextualize their expressive style: a moderate tendency toward introversion (IE $= 0.510$) implies a preference for internal processing and reduced emotional expressiveness; elevated neuroticism (NS $= 0.545$) suggests greater emotional reactivity and vulnerability to negative affect; and low judging preference (JP $= 0.283$) may indicate a less structured approach to life, potentially contributing to emotional disorganization and stress. This personality configuration supports the hypothesis that the user will likely internalize distress and express it through indirect or abstract linguistic constructs rather than explicit negative sentiment. When considered with visual features, such as solitary or symbolically reflective imagery, the user's personality traits and textual cues collectively form a coherent depressive signal. DeXMAG leverages cross-modal attention and adaptive gated fusion to align these heterogeneous but semantically related modalities, enhancing the model's capacity to detect subtle emotional indicators.

For the second sample, the proposed DeXMAG model correctly classifies the user as not depressed, despite the presence of isolated negative expressions such as "*I hate myself*". A conventional text-based model might have misclassified this instance due to its reliance on emotionally charged keywords without considering the broader context. In contrast, DeXMAG effectively integrates signals from multiple modalities to generate a more accurate prediction. Although containing specific negative phrases, the textual content primarily reflects casual remarks, humor, and expressions characteristic of informal social media discourse. The visual data comprises ordinary indoor environments, lacking any affective cues typically associated with depression. Moreover, the user's MBTI profile reveals high extroversion ($IE = 0.128$), elevated neuroticism ($NS = 0.793$), and a predominantly thinking-oriented cognitive style ($TF = 0.771$), suggesting a tendency toward emotional expressiveness rather than internalized psychological distress. By incorporating personality-aware modeling and cross-modal attention, DeXMAG can interpret depressive risk as nuanced and individualized, avoiding false-positive classifications driven by superficial textual indicators. The integration of personality-aware modeling enables DeXMAG to personalize its interpretation strategy according to the user's expressive and cognitive style, thereby facilitating more accurate identification of depression in cases where traditional models may fail.

### 4) ABLATION STUDIES

To further assess the contribution of each component within the DeXMAG model, we conducted a series of ablation studies presented in Fig. 6. We systematically evaluated the model under four configurations: (i) text-only input, (ii) image-only input, (iii) combined text and image input, and (iv) full multimodal input with personality trait integration. We selected MentalRoBERTa for textual feature extraction and VGG-16 for visual feature encoding, based on the superior performance demonstrated in the experimental results presented in Fig. 6.

The text-only configuration already provides a strong baseline, achieving an accuracy of 0.9252, a precision of 0.9550, a recall of 0.9086, and an F1-score of 0.9312. These results highlight that textual features derived from users' posts are highly informative for depression detection, capturing linguistic signals that reflect psychological states. In contrast, the image-only configuration exhibits significantly lower performance across all metrics, with an accuracy, precision, recall, and F1-score of 0.7328. This performance indicates that, in isolation, visual data contributes less discriminative power in this context, likely due to the limited emotional granularity conveyed by social media images compared to text. However, the recall of 0.7328 suggests that some depressive cues are present visually, though insufficient for reliable detection alone. When text and image modalities are combined, performance improves across all metrics, yielding an accuracy of 0.9320, precision of 0.9373, recall of 0.9319, and an F1-score of 0.9396. The increase in recall and F1-score suggests that image features serve as complementary cues, reinforcing the textual signals and helping to detect depressive patterns that might otherwise be missed.

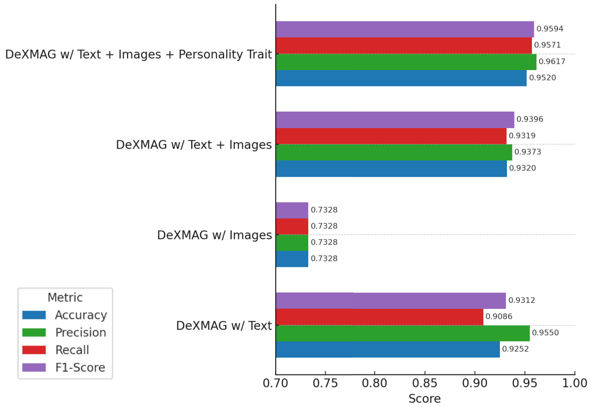


**FIGURE 6. Ablation study results evaluating the impact of individual modalities and personality trait integration on DexMAG Model performance.**

The most notable improvement occurs when MBTI personality traits are incorporated alongside text and image features. The complete model achieves the highest performance with an accuracy of 0.9520, precision of 0.9617, recall of 0.9571, and an F1-score of 0.9594. The simultaneous increase in precision and recall indicates that the model captures more true positives (higher recall) and reduces false positives (higher precision), enhancing sensitivity and specificity. These results confirm that user-level psychological profiling, in the form of personality traits, provides valuable contextual information that allows the model to make more personalized and accurate inferences. Our ablation study demonstrates that each modality contributes incrementally to overall performance. Text remains the dominant modality, while images enhance robustness when fused appropriately. Most critically, including personality traits significantly strengthens the model's predictive capacity, validating the core hypothesis of this work: that personalization through psychological features improves multimodal depression detection on social media.

## V. CONCLUSION

This study demonstrates the effectiveness of a personalized multimodal framework for depression detection by integrating textual and visual information from social media posts with user-level psychological attributes, specifically MBTI personality traits. We evaluated multiple feature extraction strategies to identify the most effective modality-specific representations. For text, transformer-based encoders such as RoBERTa and domain-specific models like MentalBERT were examined, while image features were extracted using deep convolutional networks including VGG-16 and ResNet50. The MentalRoBERTa–VGG-16 combination yielded superior performance, highlighting the importance of modality-aware extractor selection.

Furthermore, personality-aware modeling, operationalized through inferred MBTI probabilities, proved valuable in capturing user-specific affective and behavioral tendencies. The model achieved consistent improvements across all evaluation metrics when fused with multimodal content via an enhanced cross-modal attention module augmented by adaptive gated fusion. Ablation experiments confirmed the complementary nature of textual, visual, and personality-derived features. The performance comparison results demonstrated that the proposed fusion strategy was pivotal in achieving significant performance improvements compared to single-modality baselines and existing state-of-the-art fusion methods.

Ultimately, this research provides a novel contribution to the field of computational mental health by advancing a domain-adaptive and user-centric framework for automatic, personalized depression detection. Future work may focus on extending the model to capture temporal patterns, longitudinal user behavior, and social network interactions, thereby enriching the contextual understanding of individual mental health trajectories. In addition, enhancing model interpretability through post hoc explainability methods such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) will be essential for increasing transparency and trustworthiness, particularly in sensitive clinical or real-world deployment scenarios.

Beyond model development, future work should explore the real-world deployment of DeXMAG within clinically relevant and ethically governed frameworks. This includes integrating the model into opt-in digital mental health platforms where users consent to monitoring, and designing mechanisms to deliver supportive guidance while preserving privacy. Additionally, the development of responsible emergency response protocols, particularly in cases where suicidal ideation is detected, will require collaboration with clinical professionals to ensure ethical verification, minimize false positives, and maintain compliance with legal standards. These directions are essential to advance DeXMAG toward practical and trustworthy mental health applications.

## ACKNOWLEDGMENT

The authors gratefully acknowledges the financial support that enabled the successful completion of this research.

## REFERENCES

[1] A. Anshul, G. S. Pranav, M. Z. U. Rehman, and N. Kumar, "A multimodal framework for depression detection during COVID-19 via harvesting social media," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 2, pp. 2872–2888, Apr. 2024.

[2] J. J. D. Leow, H. N. Chua, M. B. Jasser, B. Issa, and R. T. K. Wong, "Comparison of depression detection between LLMs and zero-shot learning using DAD dataset," in *Proc. 21st IEEE Int. Colloq. Signal Process. Appl. (CSPA)*, Feb. 2025, pp. 295–300.

[3] M. Li, Y. Wei, Y. Zhu, S. Wei, and B. Wu, "Enhancing multimodal depression detection with intra- and inter-sample contrastive learning," *Inf. Sci.*, vol. 684, Dec. 2024, Art. no. 121282.

[4] L. Wang, Y. Zhang, B. Zhou, S. Cao, K. Hu, and Y. Tan, "Automatic depression prediction via cross-modal attention-based multi-modal fusion in social networks," *Comput. Electr. Eng.*, vol. 118, Aug. 2024, Art. no. 109413.

[5] T. Ahmed, S. Ivan, A. Munir, and S. Ahmed, "Decoding depression: Analyzing social network insights for depression severity assessment with transformers and explainable AI," *Natural Lang. Process. J.*, vol. 7, Jun. 2024, Art. no. 100079.

[6] N. K. Iyortsuun, S.-H. Kim, H.-J. Yang, S.-W. Kim, and M. Jhon, "Additive cross-modal attention network (ACMA) for depression detection based on audio and textual features," *IEEE Access*, vol. 12, pp. 20479–20489, 2024.

[7] Z. Chen, D. Wang, L. Lou, S. Zhang, X. Zhao, S. Jiang, J. Yu, and J. Xiao, "Text-guided multimodal depression detection via cross-modal feature reconstruction and decomposition," *Inf. Fusion*, vol. 117, May 2025, Art. no. 102861.

[8] K. M. Hasib, M. R. Islam, S. Sakib, M. A. Akbar, I. Razzak, and M. S. Alam, "Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 4, pp. 1568–1586, Aug. 2023.

[9] G. A. Pradnyana, W. Anggraeni, E. M. Yuniarno, and M. H. Purnomo, "An explainable ensemble model for revealing the level of depression in social media by considering personality traits and sentiment polarity pattern," *Online Social Netw. Media*, vol. 46, May 2025, Art. no. 100307.

[10] L. Ilias and D. Askounis, "Multitask learning for recognizing stress and depression in social media," *Online Social Netw. Media*, vols. 37–38, Sep. 2023, Art. no. 100270.

[11] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Classification of mental illnesses on social media using RoBERTa," in *Proc. 12th Int. Workshop Health Text Mining Inf. Anal.*, Apr. 2021, pp. 59–68.

[12] C. M. Greco, A. Simeri, A. Tagarelli, and E. Zumpano, "Transformer-based language models for mental health issues: A survey," *Pattern Recognit. Lett.*, vol. 167, pp. 204–211, Mar. 2023.

[13] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," in *Proc. 13th Lang. Resour. Eval. Conf.*, Marseille, France, Jun. 2021, pp. 7184–7190.

[14] W. B. Tahir, S. Khalid, S. Almutairi, M. Abohashrh, S. A. Memon, and J. Khan, "Depression detection in social media: A comprehensive review of machine learning and deep learning techniques," *IEEE Access*, vol. 13, pp. 12789–12818, 2025.

[15] T. Zhang, K. Yang, S. Ji, and S. Ananiadou, "Emotion fusion for mental illness detection from social media: A survey," *Inf. Fusion*, vol. 92, pp. 231–246, Apr. 2023.

[16] M. D. Choudhury and S. De, "Mental health discourse on Reddit: Self-disclosure, social support, and anonymity," in *Proc. Int. AAAI Conf. Web Social Media*, 2014, vol. 8, no. 1, pp. 71–80.

[17] U. Pavalanathan and M. De Choudhury, "Identity management and mental health discourse in social media," in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 315–321.

[18] S. Li and Y. Xiao, "A depression detection method based on multi-modal feature fusion using cross-attention," 2024, *arXiv:2407.12825*.

[19] A. Malhotra and R. Jindal, "Deep learning techniques for suicide and depression detection from online social media: A scoping review," *Appl. Soft Comput.*, vol. 130, Nov. 2022, Art. no. 109713.

[20] S. Ahmed, M. A. Yousuf, M. M. Monowar, A. Hamid, and M. O. Alassafi, "Taking all the factors we need: A multimodal depression classification with uncertainty approximation," *IEEE Access*, vol. 11, pp. 99847–99861, 2023.

[21] T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen, "Cooperative multimodal approach to depression detection in Twitter," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 110–117.

[22] M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian, "Multimodal spatiotemporal representation for automatic depression level detection," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 294–307, Jan. 2023.

[23] H. 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. 2023.

[24] 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.

[25] L. Zhou, Z. Liu, Z. Shangguan, X. Yuan, Y. Li, and B. Hu, "TAMFN: Time-aware attention multimodal fusion network for depression detection," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 669–679, 2023.

[26] J. Hong, J. Lee, D. Choi, and J. Jung, "Depression level prediction via textual and acoustic analysis," *Comput. Biol. Med.*, vol. 190, May 2025, Art. no. 110009.

[27] Y. Zhou, X. Yu, Z. Huang, F. Palati, Z. Zhao, Z. He, Y. Feng, and Y. Luo, "Multi-modal fused-attention network for depression level recognition based on enhanced audiovisual cues," *IEEE Access*, vol. 13, pp. 37913–37923, 2025.

[28] A.-M. Bucur, A. Cosma, P. Rosso, and L. P. Dinu, "It's just a matter of time: Detecting depression with time-enriched multimodal transformers," in *Proc. Eur. Conf. Inf. Retr.*, 2023, pp. 200–215.

[29] S. Yang, L. Cui, L. Wang, T. Wang, and J. You, "Enhancing multimodal depression diagnosis through representation learning and knowledge transfer," *Heliyon*, vol. 10, no. 4, Feb. 2024, Art. no. e25959.

[30] J. S. L. Figuerêdo, A. L. L. M. Maia, and R. T. Calumby, "Early depression detection in social media based on deep learning and underlying emotions," *Online Social Netw. Media*, vol. 31, Sep. 2022, Art. no. 100225.

[31] M. Fang, S. Peng, Y. Liang, C.-C. Hung, and S. Liu, "A multimodal fusion model with multi-level attention mechanism for depression detection," *Biomed. Signal Process. Control*, vol. 82, Apr. 2023, Art. no. 104561.

[32] X. Long, Y. Zhang, X. Shu, and J. Shu, "Image-text fusion model for depression tendency detection based on attention," in *Proc. 6th Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2023, pp. 730–734.

[33] Z. Li, Z. An, W. Cheng, J. Zhou, F. Zheng, and B. Hu, "MHA: A multimodal hierarchical attention model for depression detection in social media," *Health Inf. Sci. Syst.*, vol. 11, no. 1, p. 6, Jan. 2023.

[34] W. Kang, F. Steffens, S. Pineda, K. Widuch, and A. Malvaso, "Personality traits and dimensions of mental health," *Sci. Rep.*, vol. 13, no. 1, p. 7091, May 2023.

[35] Y. Li, D. Wei, J. Sun, J. Meng, Z. Ren, L. He, K. Zhuang, and J. Qiu, "Personality and depression: A review of theory model and behavior and neural mechanism," *Acta Physiolog. Sinica*, vol. 71, no. 1, pp. 163–172, 2019.

[36] P. Boyce, G. Parker, B. Barnett, M. Cooney, and F. Smith, "Personality as a vulnerability factor to depression," *Brit. J. Psychiatry*, vol. 159, no. 1, pp. 106–114, Jul. 1991.

[37] Y. Takahashi, B. W. Roberts, S. Yamagata, and N. Kijima, "Personality traits show differential relations with anxiety and depression in a nonclinical sample," *Psychologia*, vol. 58, no. 1, pp. 15–26, 2015.

[38] D. Zhang, E. S. Huebner, and L. Tian, "Longitudinal associations among neuroticism, depression, and cyberbullying in early adolescents," *Comput. Hum. Behav.*, vol. 112, Nov. 2020, Art. no. 106475.

[39] W. Liying and S. Sheibani, "The relationship between the myers–briggs type indicator (MBTI) types and psychological well-being among college students in China," *J. Ecohumanism*, vol. 3, no. 7, pp. 3611–3619, Oct. 2024.

[40] J. P. Thekkekara, S. Yongchareon, and V. Liesaputra, "An attention-based CNN-BiLSTM model for depression detection on social media text," *Expert Syst. Appl.*, vol. 249, Sep. 2024, Art. no. 123834.

[41] M. Kerasiotis, L. Ilias, and D. Askounis, "Depression detection in social media posts using transformer-based models and auxiliary features," *Social Netw. Anal. Mining*, vol. 14, no. 1, p. 196, Sep. 2024.

[42] R. Poświata and M. Perełkiewicz, "OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models," in *Proc. 2nd Workshop Lang. Technol. Equality, Diversity Inclusion*, May 2022, pp. 276–282.

[43] F. A. Nazira, S. R. Das, S. A. Shanto, and M. F. Mridha, "Depression detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Signal Process., Inf., Commun. Syst. (SPICSCON)*, Dec. 2021, pp. 9–13.

[44] J. Cha, S. Kim, D. Kim, and E. Park, "MOGAM: A multimodal object-oriented graph attention model for depression detection," 2024, *arXiv:2403.15485*.

[45] J. Wei and J. Zhou, "Exploring personality and emotion in risk prediction of depression," in *Proc. IEEE/ACIS 27th Int. Conf. Softw. Eng., Artif. Intell., Netw. Parallel/Distrib. Comput. (SNPD)*, Jul. 2024, pp. 228–233.

[46] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[47] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2018, pp. 4171–4186.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[50] G. A. Pradnyana, W. Anggraeni, E. M. Yuniarno, and M. H. Purnomo, "Enhancing MBTI personality trait prediction from imbalanced social media data using hybrid query expansion ranking and Glo Ve-BiLSTM," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ)*, Aug. 2023, pp. 1–6.

[51] A. Bruno and G. Singh, "Personality traits prediction from text via machine learning," in *Proc. IEEE World Conf. Appl. Intell. Comput. (AIC)*, Jun. 2022, pp. 588–594.

[52] R. C. Pradana and D. Suhartono, "Synonym replacement augmentation for handling data imbalance in personality classification," in *Proc. 8th Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, Aug. 2024, pp. 1–6.

[53] M. An, J. Wang, S. Li, and G. Zhou, "Multimodal topic-enriched auxiliary learning for depression detection," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, Dec. 2020, pp. 1078–1089.

**GEDE ADITRA PRADNYANA** (Graduate Student Member, IEEE) received the bachelor's degree in computer science from Universitas Udayana, Indonesia, in 2011, and the master's degree in informatics engineering from the Institut Teknologi Sepuluh Nopember (ITS), Indonesia, where he is currently pursuing the Ph.D. degree with the Department of Electrical Engineering. He also lectures with Universitas Pendidikan Ganesha, Indonesia. His research interests primarily include natural language processing, text mining, deep learning, and data science.

**WIWIK ANGGRAENI** (Member, IEEE) received the Bachelor of Science degree in mathematics from the Institut Teknologi Sepuluh Nopember (ITS), Indonesia, in 1997, the Master of Computer Science degree from the Department of Informatics Engineering, Faculty of Information Technology, ITS, in 2003, and the Ph.D. degree in electrical engineering from the Faculty of Intelligent Electrical and Informatics Technology (FTEIC), ITS, in 2022. She is currently a Professor of predictive analytics in big data with the Department of Information Systems, ITS. Her research interests include predictive analytics, forecasting, and data mining.

**EKO MULYANTO YUNIARNO** (Member, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in electrical engineering from the Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 1995, 2005, and 2013, respectively. He is currently a Senior Lecturer with the Department of Computer Engineering, ITS. His research interests include computer vision, image processing, machine learning, and deep learning.

**MAURIDHI HERY PURNOMO** (Senior Member, IEEE) received the Ph.D. degree from Osaka City University, Osaka, Japan, in 1995. He is currently a Professor with the Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, where he is actively involved in teaching and research in the areas of research methodology, artificial intelligence, neural networks, and image processing. His research interests include smart grids, renewable energy, artificial intelligence applications in healthcare, and power systems. He currently serves as the Chair for the IEEE Industrial Electronics Society, Indonesia Section.