International Research Journal of Multidisciplinary Scope (IRJMS), 2025; 6(3):1587-1610

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2025.v06i03.04620

FMindMonitorAI: An AI-Driven Framework for Social Media-Based Mental Health Analysis Using MindFusionNet

Srinivas Kanakala[1]*, Vempaty Prashanthi[2], Veluru Chinnaiah[3], KV Sharada[4], M Sumalatha[5]

[1]Department of CSE, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India, [2]Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad, India, [3]Department of CSE, Vijaya Engineering College, Khammam, Telangana, India, [4]Department of CSE, Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana, India, [5]Department of CSE, CVR College of Engineering, Hyderabad, India. *Corresponding Author's Email: srinivaskanakala@gmail.com

Abstract

Mental health disorders such as depression, anxiety, and stress are rising globally, necessitating effective and scalable monitoring solutions. Social media platforms provide an abundant source of real-time behavioral and linguistic data, which can be leveraged for early detection of mental health conditions. However, most existing methodologies rely solely on textual features and traditional machine learning or deep learning models, limiting their ability to capture the multi-faceted nature of user behavior and interactions. These approaches often lack interpretability and fail to generalize well across diverse social contexts, reducing their practical applicability. To address these limitations, this paper proposes MindMonitorAI, an AI-driven framework for mental health analysis using social media data. The framework introduces a novel hybrid deep learning model, MindFusionNet, which integrates three complementary modalities: textual posts, behavioral engagement metrics, and social graph connections. The architecture employs BERT-based encoders for text analysis, neural networks for behavioral features, and graph neural networks for modeling social interactions, followed by attention-based multi-modal feature fusion. Explainability techniques, including SHAP and LIME, are incorporated to enhance transparency and support ethical decision-making. An experimental evaluation of multiple publicly available mental health datasets demonstrates that MindFusionNet achieves superior performance, attaining an accuracy of 98.68%, outperforming baseline machine learning and deep learning models. The results validate the efficacy of multi-modal integration and attention mechanisms in improving predictive capabilities. The proposed framework is significant for real-time, interpretable mental health monitoring systems, providing actionable insights for healthcare professionals and supporting timely intervention strategies.

Keywords: Behavioral Analysis, Explainable AI, Mental Health Analysis, Multi-Modal Deep Learning, Social Media Monitoring.

Introduction

Mental health disorders such as depression, anxiety, and stress are becoming increasingly more widespread and serious global problems as our modern lifestyle increasingly changes, and we see demands for more significant digital interaction. Social media platforms provide a wealth of data reflecting users in close to real-time, which can be an invaluable resource for understanding their thoughts feelings, and actions. Many studies have shown the feasibility of using machine learning and deep learning techniques to analyze social media data to identify mental health disorders (1, 2). Most current methods, however, are heavily geared towards textual content, often ignoring behavioral habits and social interaction patterns, which can provide further insight into a user's mental state. However, some key limitations of current methodologies have also been identified in recent literature. Although traditional machine learning models, such as Support Vector Machines (SVM) and Random Forests, have reached particular effectiveness, they are limited by feature engineering and low generalization across different social contexts (3). Likewise, deep learning models (CNNs or recurrent architectures) typically input only textual features, while overlooking critical behavioral indicators and social network linkages (4, 5). Additionally, the inability to interpret most models makes it difficult to translate them into clinical or real-world scenarios.

(Received 24th March 2025; Accepted 09th July 2025; Published 30th July 2025)

The association between language usage and mental health has a strong theoretical foundation in psychology. According to cognitive-behavioral theory, those with depression or anxiety often show negative, self-focused language and cognitive distortion. The LIWC platform has even been employed to determine the presence of markers like first-person singular pronouns, negative emotion words, or absolutist terms in text that are associated with depressive states. Adopting such linguistic features into the AI model enables it to learn more precise and theoretically motivated assessments of mental health.

To bridge these gaps, this paper presents MindMonitorAI, a novel AI-based framework for monitoring mental health through social media activity. The proposed work aims to develop a deep learning model called MindFusionNet, which effectively integrates and utilizes multimodal data consisting of textual posts, behavioral engagement metrics, and social graph interactions to improve prediction accuracy while yielding interpretable insights. The research innovation uses attention-based multi-modal fusion and explainability techniques, yielding high performance and interpretability. Our research contributions are: (a) a novel hybrid framework, the combination of BERT-based textual encoders, neural behavioral encoders, and graph neural networks to mine social features; (b) a carefully designed attention mechanism that optimally fuses multi-modal features; (c) thorough evaluation on various datasets; and (d) transparent model explanations using SHAP and LIME.

The remainder of the paper is structured as follows: Section 2 presents a systematic review of relevant literature, summarizing existing machine learning and deep learning techniques for mental health analysis. Section 3 details the proposed MindMonitorAI framework and MindFusionNet model, elaborating on the data modalities, feature extraction, model architecture, and training strategies. Section 4 describes the experimental setup, dataset description, performance evaluation, and comparative analysis with existing models. Section 5 discusses the findings, highlighting how the study addresses limitations in prior works and outlines the specific limitations of the current research. Finally, Section 6 concludes the paper and suggests future research directions to enhance scalability, real-time deployment, and clinical applicability.

Existing studies employ machine learning and deep learning models for mental health detection, often relying solely on textual data from social media. The potential applications of machine learning in the healthcare industry have been explored, with a focus on early epidemic detection, personalized treatment, and operational efficiency. The role of machine learning in reducing healthcare costs while improving outcomes was highlighted (1). A comparative analysis was conducted between virtual influencers (VIs) and human influencers, addressing their emergence, ethical challenges, and marketing applications, while also stressing the need for further advancements in AI (2). The application of artificial intelligence in mental health care has been extensively investigated, with attention given to technological advancements, ethical considerations, and legal implications. The necessity for transparency, validation, and sustained research has been emphasized to ensure the effective implementation of this approach (3). The convergence of AI and social media has been analyzed, focusing on developments in content filtering and mental health detection, along with associated ethical concerns and recommendations for future AI-human collaboration (4). Stakeholder collaboration in mental health AI research has been highlighted, with the importance of involving patients as subject matter experts and fostering consensus to support human-centered AI being stressed (5).

Concerns about AI, as expressed through expert opinion and social media analysis, have been studied, leading to the identification of seven critical challenges. Solutions emphasizing collaboration and regulatory mechanisms were proposed for the ethical integration of AI (6). A sound-based therapy system driven by AI has been introduced to support the mental well-being of older adults, with preliminary results suggesting a reduction in loneliness. However, further integration with other technologies remains necessary (7). The potential of AI in prenatal mental health has been examined, with a focus on applications such as voice assistants and risk prediction systems. Issues concerning interpretability, class imbalance, and data quality were acknowledged (8). Machine learning

techniques for detecting mental health conditions in online social networks (OSNs) have been evaluated, demonstrating promise while highlighting limitations related to algorithmic innovation and data reliability (9). Facial expression recognition and sentiment analysis methods for detecting mental health disorders on social media platforms have been evaluated, with HSV color-based tone prediction applied to visual content (10).

A study was conducted to classify mental health disorders among Twitter users communicating in both English and Spanish, utilizing traditional and deep learning techniques. The approach demonstrated strong classification performance across multilingual datasets (11). Artificial intelligence models were reviewed for mental health analysis on social media, focusing on key components such as feature extraction, available datasets, suicide detection, and identified gaps for future exploration (12). Natural language processing and machine learning methods were analyzed for detecting mental health conditions in social media contexts, with a specific focus on challenges, weaknesses, and potential enhancements in predictive modelling (13). A machine learning strategy was proposed for depression detection that remains effective even when applied to unrelated datasets, with future directions pointing toward the use of unsupervised learning approaches (14). Depression and anxiety identification from social media posts was explored using machine learning models, emphasizing the significance of early detection in post-COVID-19 scenarios while also noting ethical limitations (15).

A framework based on LSTM was introduced for the early detection of mental illness from Twitter data, achieving performance improvements over existing techniques and suggesting broader applicability in monitoring social behaviors (16). A method combining machine learning and BERT was described for detecting stress in Twitter users, with future research focused on accuracy enhancement and thematic analysis of social issues (17). Approaches involving artificial intelligence and machine learning for diagnosing depression were evaluated, incorporating sentiment analysis and emotion recognition, with plans to integrate these methods into clinical decision-making systems in the future (18). An explainable deep learning model, MDHAN, was presented for depression detection on social media, demonstrating superior performance by combining interpretability with predictive power. Future work aims to expand this by integrating additional data sources (19). Sentiment analysis techniques were applied to assess mental health concerns during the COVID-19 pandemic, with planned extensions addressing ethical considerations and support for multilingual content (20).

The application of natural language processing and sentiment analysis for detecting mental health conditions was explored, with tools such as Python and NLTK highlighted for their potential in identifying depression and cyberbullying cases (21). Topic and sentiment analysis techniques were employed to evaluate 104 mental health applications, revealing both positive and negative themes and suggesting improvements for future app design and reviews (22). Deep learning models were utilized to identify mental health disorders—including self-harm, anorexia, and depression—based on social media language patterns, with a focus on emotional and linguistic indicators (23). A comprehensive review of sentiment analysis methods and datasets was conducted, identifying recent advancements and highlighting unresolved challenges that future frameworks aim to address (24). Mental health classification using deep learning, multiple algorithms, and Reddit data was investigated, with future enhancements targeting deeper demographic representation and model improvement (25).

An unsupervised learning model was proposed to detect early signs of mental illness among Twitter users, with future validation to be conducted in collaboration with medical professionals (26). A framework called SocialText was introduced to extract linguistic features related to personality, loneliness, and anxiety from private Facebook messages, with future work directed toward deeper DTC (disorder-to-cause) analysis (27). Transformer-based architectures such as BERT and RoBERTa were recommended to improve the accuracy of detecting non-medical prescription drug use from Twitter content, with further development intended to enhance model performance (28). A personalized prediction model for psychological constructs such as anxiety, personality, and sleep quality was developed using

multimodal data from sensors and social media, providing a foundation for emotion-aware computing (29). A stress prediction technique was introduced, integrating personality traits, anticipated life events, and memory networks using social media posts. However, the limited availability of high-quality data remains a significant constraint (30).

Transformer-based models were proposed for the detection of suicidal ideation from social media data, showing improved performance over Bi-LSTM architectures. However, the study noted limitations, including annotation bias and a small dataset size, with future work planned to focus on data expansion and refinement (31). Natural language processing and machine learning techniques were applied to accurately identify sadness in Arabic-language social media posts, with future directions targeting the classification of depression and enhancements in dataset collection (32). A method leveraging social media data was introduced to assess the risk of human interaction and perceptions of lockdown, with findings showing a moderate correlation with social alienation, despite limitations in data quality and practical applicability (33). An F-measure-based approach was proposed to analyze user profiles, including gender, age, and personality traits, using natural language processing, fuzzy logic, and semantic enrichment. Cross-linguistic validation was identified as a key area for future research (34). Deep learning and NLP methods were evaluated for classifying healthcare-related texts on social media, demonstrating that hybrid models outperform individual approaches, and suggesting that further refinement is necessary to enhance accuracy (35).

A comprehensive mental health monitoring system was tested using facial recognition, wearable bright clothing, and cloud terminals integrated with robotics, offering a novel framework for emotional support. Future improvements are expected in usability and deployment (36). A model combining AI and NLP was developed to accurately predict cognitive disorders, offering a foundation for future advancements in psychiatric healthcare systems (37). An LSTM-based IoT framework was introduced for emotion detection, demonstrating low latency, high reliability, and firm performance in remote learning and healthcare applications during the pandemic (38). A study was conducted with 272 participants to understand expectations of AI assistant proactivity, yielding design insights to enhance the adoption and responsiveness of intelligent systems (39). An AI-driven 5G+ wireless network personalization framework was presented to optimize user satisfaction and resource utilization, showing increased efficiency while acknowledging areas for continued optimization (40).

The literature reveals that traditional models often overlook behavioral and social graph features, which limits their predictive performance and interpretability. Most methods lack multi-modal fusion and transparency. This gap underscores the need for an integrated, explainable framework. MindMonitorAI addresses these limitations by incorporating multi-modal data and advanced attention-based fusion strategies.

Methodology

Proposed Framework

This section presents the proposed MindMonitorAI framework for analyzing mental health based on social media. The framework combines features from multimodal inputs, including text posts, user behavioral data, and social graphs. It proposes a fundamental label classification model called MindFusionNet. It utilizes a deep learning-based encoder for input processing and an attention-based feature fusion mechanism to achieve high prediction accuracy and interpretability.

Outline of MindMonitorAI Framework: As proposed in this paper, we introduce the MindMonitorAI framework (Figure 1), which provides an efficacious and automatic approach to monitoring and analyzing mental health using social media data. This interweaves multi-modal data-input methodologies, deep learning-based modeling, and interpretability mechanisms to elucidate subtle indicators of mental well-being. The first step in the framework involves gathering social media data from various publicly accessible datasets, namely the Reddit Mental Health Dataset, eRisk Dataset, and CLPsych Dataset. For this work, you study 3 significant social media datasets consisting of user-generated posts, comments, and metadata related to various mental health issues like depression, anxiety, stress, and self-harm risks.

After collecting the data, the framework executes a complete preprocessing pipeline for data cleaning and analysis preparation. This process involves removing noise (URLs, memorable characters, and stop words), tokenization, lemmatization, and normalization. Moreover, non-English posts were detected and standardized to ensure linguistic uniformity in the dataset. User anonymity is maintained, and sensitive information is anonymized to comply with privacy regulations, making this one of the stages that equally satisfies ethics considerations.

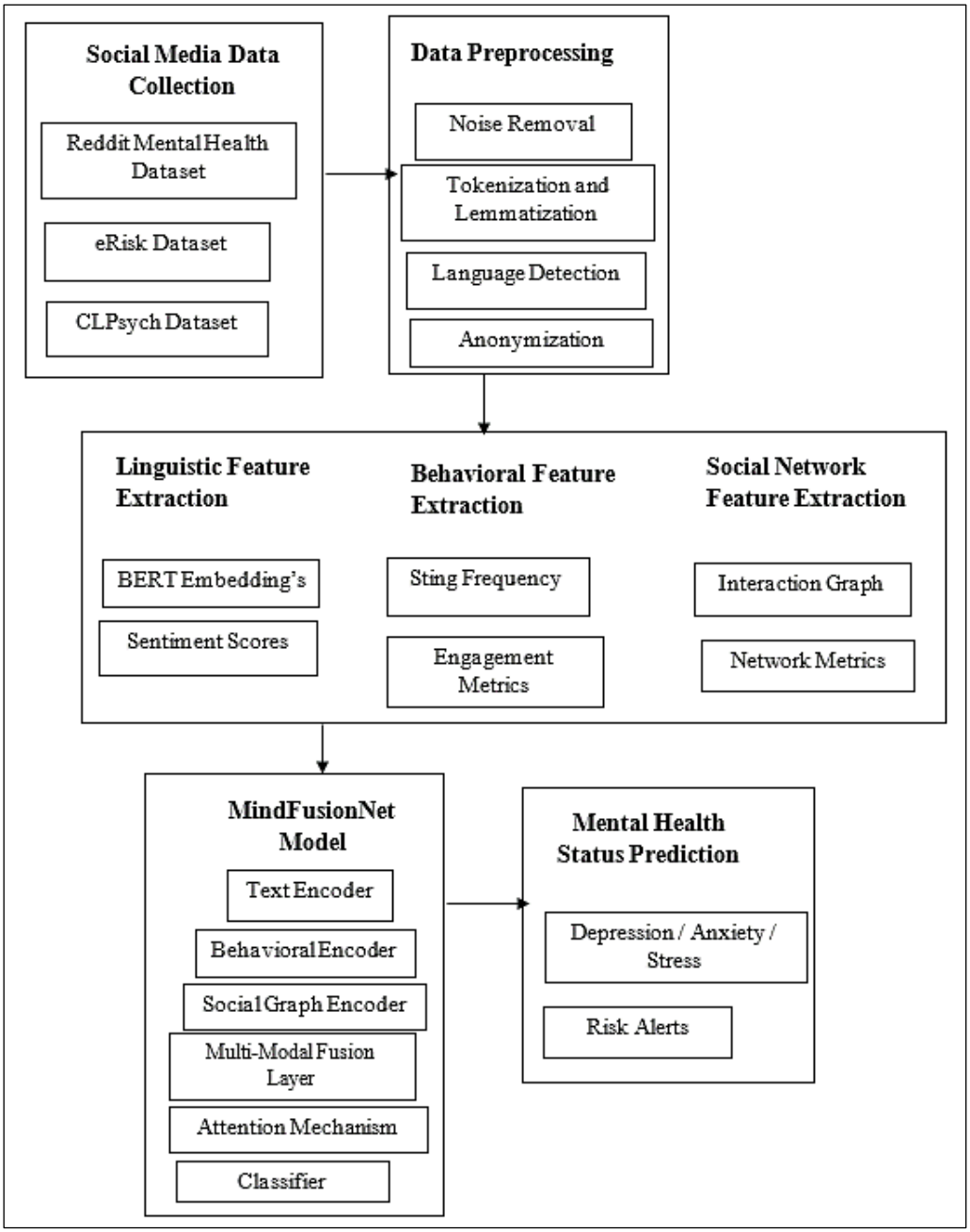

Figure 1: Overview of the MindMonitorAI Framework

After preprocessing, it proceeds to feature extraction, where MindMonitorAI extracts three main groups of features: linguistic (what users discuss), behavioral (what users do), and social network features (how users interact with others). We extract linguistic features from the textual contents of posts using a language model (BERT-based) encoder to obtain contextualized embeddings. Behavioral features such as post frequency, engagement metrics, and activity patterns are extracted to study user behavior. Social network features are encapsulated by

building interaction graphs (posts, replies, mentions, follows) where users are nodes and interaction between them is represented by edges. All those multi-modal key features are explored through the proposed deep learning model MindFusionNet, which forms the foundation of the entire framework. It uses three separate encoders: a Transformer-based text encoder for linguistic information, a feedforward NN for behavioral data, and a GNN for social network data. The outputs of these encoders are merged into a single feature representation, which is then enhanced with attention mechanisms. The second portion of the model is the classification side, which predicts the users' mental health status by classifying them into the most appropriate mental conditions.

To make the framework more transparent and more trusted, explainability modules (SHAP, LIME) are integrated. These modules yield explanatory insights that highlight the features and behaviors driving the model's predictions. Finally, MindMonitorAI produces two components as its output: a prediction of the mental health state, along with an explanation of the prediction in visual form. This is in line with the intention of this study, which caters to the clinical needs of healthcare professionals and social media moderators, who are ultimately the decision-makers for patients.

Data Collection and Preprocessing

The data used in the MindMonitorAI framework is derived from three publicly available social media datasets (the Reddit Mental Health Dataset, the eRisk Dataset, and the CLPsych Dataset). The datasets include a large amount of user-generated content (e.g., posts, comments, and metadata) related to mental health topics such as depression, anxiety, stress, and self-harm. The Reddit Mental Health Dataset is compiled from posts from various mental health subreddits (e.g., r/depression, r/anxiety) known to be used in earlier studies as indicators of mental health status through self-disclosure. Although there was no empirical basis, these subreddit postings can be considered self-report proxies for mental health tagging. CLPsych Dataset: Annotated social media post data with mental health categories. Supervised learning source.

A highly complex preprocessing pipeline then processes the raw data to ensure data quality is controlled and consistent. This also includes removing irrelevant noise, such as URLs, HTML tags, memorable characters, and excessive white space. Meanwhile, bot-like behavior was removed by filtering words that have been used repeatedly in posts or posted too frequently in a timeframe. Spam content is detected through keyword-based heuristics and an account-level activity threshold. Informal and internet words were normalized using a custom mapping of slang to standard words. Through these pre-processing steps, the system keeps only human-like, context-rich posts for feature extraction.

All data was anonymised and cleaned to safeguard the privacy of users and for ethical reasons. While usernames, IDs, and other explicit identifiers were deleted. All external hyperlinks, their timestamps, and geolocation information were removed or altered. Furthermore, Dadooth and Sketches were used only for publicly available data of the consented forums (such as the Reddit subreddits) to be consistent with ethical practices. This anonymization guarantees that individual identities cannot be reconstructed from the processed information. All notations used in the proposed framework are given in Table 1.

Table 1: Notations Used in MindMonitorAI Framework

Notation	Description
T_i	i[th] token in a social media post
n	Total number of tokens in a post
E_{text}	Textual embedding vector generated by BERT encoder
P_f	Normalized posting frequency feature
T_p	The average time interval between consecutive posts
E_m	Vector of engagement metrics (likes, shares, comments)
E_{behav}	Behavioral feature vector comprising P_f, T_p, E_m
G=(V,E)	Social interaction graph with nodes VV (users) and edges EE (interactions)

N(v)	Set of neighboring nodes for user node v
$E_{graph}^{(k)}(v)$	Embedding of node v at layer kk in the GNN encoder
$W^{(k)}, b^{(k)}$	Weight matrix and bias in the kthk^{th} layer of the GNN encoder
σ	Nonlinear activation function (ReLU)
E_{fusion}	Concatenated multi-modal feature vector (E_{text}),
W_{fusion}, b_{fusion}	Weight matrix and bias in the fusion dense layer
H_{fusion}	Hidden feature representation after fusion dense layer
$W_{attn,}b_{attn}, v$	Learnable parameters in the attention mechanism
α_i	Attention weight for the ithi^{th} feature in the fused vector
H_{attn}	Attention-refined feature vector
Z_c	Logit score for class cc in classification layer
$\hat{y}_c$	Predicted probability for class cc
C	Total number of mental health condition classes
y_c	Ground truth one-hot encoded label for class cc
L	Cross-entropy loss function
$\emptyset_i$	SHAP value representing contribution of feature ii
S	Subset of features excluding feature ii
F	Set of all features
f(S)	Model output when only features in subset SS are considered

Feature Extraction: Feature extraction is a crucial process within the MindMonitorAI framework, highlighting the diverse representational demands of social media data for mental health analysis. The information gathered from Reddit, eRisk, and CLPsych datasets varies, including textual content, behavior patterns, and social network interactions.

$$E_{text} = BERT(|T_1, T_2, \dots, T_n|)$$

Where n is the number of tokens in the post. It captures aspects of semantics and syntax, helping the model learn contextual dependencies and sentiment-infused expressions indicative of a mental health condition. Behavioral features and textual data are generated based on user activity and engagement. Such as frequency of posts,

$$E_{behav} = [P_f, T_p, E_m]$$

While many of these features can give valuable indications, such as sharp increases in the frequency of posting or a lack of responses that may indicate changes in mental health status, they do not, by themselves, offer any indication of functioning. In addition, social interaction patterns are modeled by building user interaction graphs.

To obtain the linguistic subtly, we tokenize each social media post and then encode it with a pre-trained BERT model, which generates contextualized embeddings for the words. Let be T_i the i^{th} token that one can put in a post. The BERT encoder encodes a sequence of tokens to a dense vector representation E_{text} as in Eq. [1].

[1]

average time interval of posts, and interaction metrics (Likes, Shares, Comments, etc.). Let P_f be the normalized posting frequency, $E_m$the engagement metric vector, and T_p The average time interval between posts. The vector of behavioral features x is formulated as in Eq. [2].

[2]

Users are modeled as nodes, and interactions (replies, advances, algorithms, and follows) are modeled as edges. G= (V, E) is a graph where the nodes of our users and the edges of E interact. We apply a graph neural network (GNN) to encode the structural information to the embedding. E_{graph} of each user node is updated through messages passing from neighboring nodes as in Eq. [3].

$$E_{graph}^{(k)}(v) = \sigma\left(\sum_{u \in N(v)} W^{(k)} E_{graph}^{(k-1)}(u) + b^{(k)}\right) \quad [3]$$

Here, n(v) is the set of neighbors of user v, $W^{(k)}$, $b^{(k)}$ And σ is a nonlinear activation function.
Finally, the step combines the extracted feature vectors. E_{text}, E_{behav}, and E_{graph} into a unified multi-modal feature vector E_{fusion}, which serves as input to the following layers of the MindFusionNet model as in Eq. [4].

$$E_{fusion} = [E_{text} \parallel E_{behav} \parallel E_{graph}] \quad [4]$$

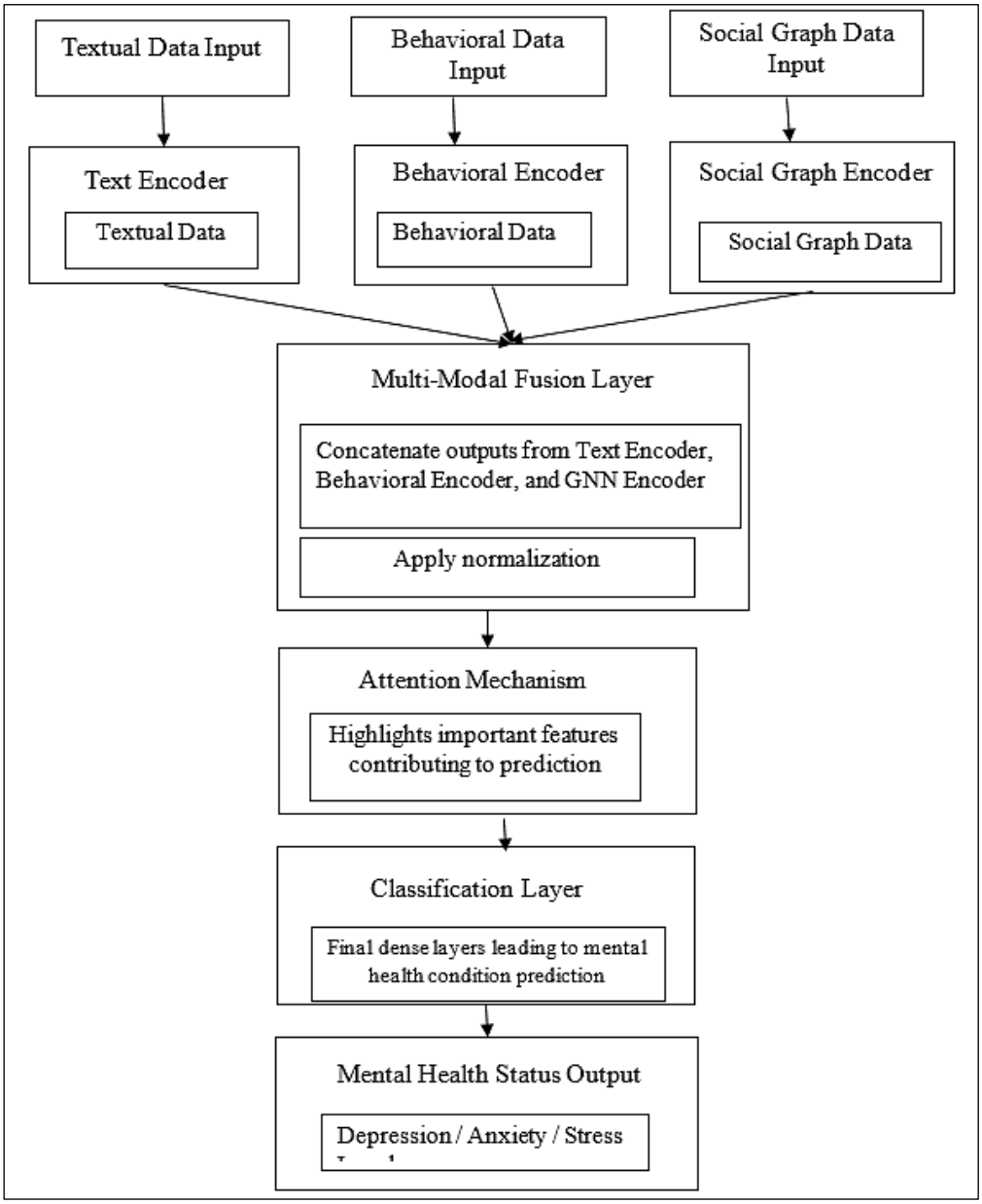

Figure 2: Architecture of MindFusionNet Model Integrating Text, Behavioral, and Social Graph Encoders with Multi-Modal Fusion and Classification Layers

By capturing language style and usage patterns, data interaction metrics, and social network analysis attributes, this extensive feature extraction process provides a prosperous and cohesive representation of linguistic, behavioral, and social dimensions, thus enabling the framework to identify nuanced trends and indicators pertinent to mental health. Behavioral attributes include posting behavior, inter-activity measurements, and temporal characteristics like time intervals between a pair of consecutive posts and time-of-day posting pattern. These time-domain features are important to identify sudden changes in stimuli/physical activity levels or circadian rhythm disturbances, often indicative of mental health conditions. They are extracted by using a feedforward neural network and are used in the multi-modal feature fusion directly.

Proposed Deep Learning Model: MindFusionNet

The deep learning model proposed in this study is MindFusionNet. It will be an essential component of the MindMonitorAI framework, which will model the multimodal features extracted from social media data to predict mental health conditions. The MindFusionNet architecture accommodates three different types of input: text-based data, behavioral data, and social interaction graphs. All these inputs are processed through separate encoders to capture relevant features before being fused for co-analysis.The first branch of the model processes the linguistic features extracted from social media posts and comments (Figure 2). A deep contextual embedding is generated using a pre-trained BERT-based transformer encoder, which creates the textual embedding vector. E_{text} given an input token sequence $[T_1, T_2, \dots, T_n]$ as in Eq. 5.

$$E_{text} = BERT(|T_1, T_2, \dots, T_n|) \quad [5]$$

This embedding preserves semantic and syntactic properties, enabling the model to identify linguistic features associated with different mental health types. Behavior features like posting frequency, engagement metrics, and temporal activity patterns are processed using a feedforward neural network parallel to the textual branch. Let denote E_{behav}The behavioral feature vector is as in Eq. [6].

$$E_{behav} = [P_f, T_p, E_m] \quad [6]$$

Where P_f represents normalized frequency, T_p the average time gap between posts, E_m and engagement measures. The behavioral encoder is transformed via several dense layers to learn a meaningful latent representation. The Third Pillar is a social interaction sequencing. We take a graph-based approach and encode users as nodes and their interactions (replies, mentions, follows) with others as edges. Let $G = (V, E)$ be the interaction graph, where the set of user nodes is represented as V The set of edges is represented as E. This GNN updates each node's embedding according to its neighbors. We compute the embedding of the user node. v layer k as in Eq. [7].

$$E_{graph}^{(k)}(v) = \sigma\left(\sum_{u \in N(v)} W^{(k)} E_{graph}^{(k-1)}(u) + b^{(k)}\right) \quad [7]$$

where $N(v)$ is the neighborhood of node v, $W^{(k)}$ and$b^{(k)}$ are weight matrix and bias, and σ is a nonlinear activation function such as ReLU. After encoding each modality, we concatenate the outputs of the three branches. E_{text}, E_{behav}, and E_{graph} into a single feature vector as in Eq. [8].

$$E_{fusion} = [E_{text} \parallel E_{behav} \parallel E_{graph}] \quad [8]$$

The concatenated vector is forwarded through a fully connected dense layer that outputs the intermediate representation to harmonize the fused features and enable effective joint learning H_{fusion} as in Eq. [9].

$$H_{fusion} = \sigma\left(W_{fusion} E_{fusion} + b_{fusion}\right) \quad [9]$$

where $W_{fusion}$and b_{fusion} are trainable parameters. We then, an attention mechanism is applied to highlight the most informative parts of the fused representation. The attention scores α It gets computed as in Eq. [10].

$$\alpha_i = \frac{expexp\,(e_i)}{\sum_{j-1}^{n} expexp\,(e_j)}\ where\ \ e_i = v^T tanh\,(W_{attn}H_{fusion} + b_{attn}) \qquad [10]$$

The feature vector H_{attn} attended with is obtained by computing the weighted sum of the components of H_{fusion} as in Eq. [11].

$$H_{attn} = \sum_{i-1}^{n}\ \alpha_i H_{fusion,i} \qquad [11]$$

Finally, the attended representation is fed through a dense classification layer, outputting a probability distribution over mental health classes with the softmax activation function as in Eq. [12].

$$\hat{y}_c = \frac{expexp\,(z_c)}{\sum_{k-1}^{c} exp\,(z_k)} \quad \forall c \in \{1,2,..,C\} \qquad [12]$$

C is the total count of mental health categories (Depression, Anxiety, Stress, etc.), and z_c is the logit score for the class c.

Proposed Algorithm

The proposed algorithm is a multimodal deep learning approach designed for mental health prediction using social media data. It integrates textual, behavioral, and social network features through specialized encoders—BERT, neural network, and GNN, respectively. The fused representation undergoes attention-based refinement, enabling accurate classification of mental health conditions while effectively capturing diverse patterns present in user behavior and content.

Algorithm 1: MindFusionNet Training Algorithm

Algorithm: MindFusionNet Training Algorithm
Input:
Preprocessed textual data D_{text},behavioral feature data D_{behav}, social graph data G = (V, E), ground truth labels Y
Output:
Trained MindFusionNet model
Steps:

- Initialize BERT encoder parameters.
- Initialize Feedforward Neural Network (FNN) parameters for behavioral features.
- Initialize Graph Neural Network (GNN) parameters for social graph features.
- For each training epoch, repeat: a. For each batch:
  Encode textual input D_{text}using BERT → E_{text}
  Encode behavioral input D_{behav}using FNN → E_{behav}
  Encode graph GG using GNN → E_{graph}
  Concatenate features: $E_{fusion} = \left[E_{text} \parallel E_{behav} \parallel E_{graph}\right]$
  Apply dense layer: $H_{fusion} = \sigma\,(W_{fusion}E_{fusion} + b_{fusion})$
  Compute attention weights $\alpha_i$and generate attended vector H_{attn}
  Pass H_{attn}through classification layer → Predicted labels $\hat{y}$
  Calculate cross-entropy loss L using Y and $\hat{y}$
  Update model parameters using Adam optimizer to minimize L
- Evaluate model performance on validation set.
- Return trained MindFusionNet model.

The training process of the MindFusionNet model is shown in Algorithm 1, where multi-modal features from social media data are obtained and used for accurate mental health prediction. The model starts by initializing the parameters of the three main components: the text encoder based on BERT, the feedforward network used for behavioral features, and the GNN used to process the graph of social interactions. The input to the model is segmented in mini-batches across the epoch for practical computation and stable convergence.

For each batch, the text first passes through the BERT encoder to obtain contextualized

embeddings that capture the nuances of the language. At the same time, behavioral characteristics, including posting frequency, engagement metrics, and temporal intervals, are passed through the feedforward neural network to obtain behavioral representations. Simultaneously, the social interaction graph is encoded through the GNN, with each user node's embedding updated iteratively based on interactions with its related nodes. Concatenated multi-modal features: The outputs of the three encoders, textual, behavioral, and graph embeddings, are concatenated directly to build a unified multi-modal feature vector.

We apply a dense layer to this concatenated feature vector to synchronize and map the features into a common latent space. The next step is to utilize an attention mechanism that assigns weights to features based on their informativeness, thereby focusing on the most informative ones. The processed and attended feature representation is finally input into a dense classification layer, which generates probability distributions over the classes of mental health conditions.

We train the model using the cross-entropy loss function, which measures the discrepancy between the predicted probabilities and the ground truth labels. Model parameters are updated using the Adam optimizer, with gradients computed by backpropagation to minimize the loss. This process is repeated over epochs until convergence. At every epoch, validation is performed to assess the model's performance, which aids in generalization and prevents overfitting. After training, the optimized MindFusionNet model is tested on the test set to examine its predictive power.

$$L = -\sum_{c-1}^{c} y_c log\ (\hat{y}_c) \quad [13]$$

Where y_c is the one hot encoded ground truth label for class c, $\hat{y}_c$ is the predicted probability for class c, and C is the total number of classes. We use the Adam optimizer because it uses adaptive learning rates, and we perform a grid search to tune the initial learning rate and weight decay hyperparameters.

Dropout regularization is employed after the dense layers to prevent overfitting, and early stopping is monitored based on validation loss. Furthermore,

Training and Evaluation Strategy

The training and evaluation strategy we adopt for the MindFusionNet model is comprehensive, ensuring robustness, generalizability, and high predictive power across various social media datasets. Using an 80:10:10 split ratio, we partition the preprocessed and feature-extracted dataset into training, validation, and testing subsets. The stratified split ensures that all class labels are represented in each subset, considering the imbalanced nature of almost all mental health datasets.

Model parameters are trained using the cross-entropy loss function, as defined in Eq. [13].

k-fold cross-validation (k=5) is performed to guarantee that the model's performance is consistent and not biased towards any single data partition.

We evaluate the model with various metrics, including Accuracy, Precision, Recall, F1-score, and Receiver Operating Characteristic - Area Under Curve (ROC-AUC). Accuracy is computed as in Eq. [14].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad [14]$$

TN, FP, and FN mean true positive, true negative, false positive, and false negative. Precision and Recall are defined as in Eq. [15]. Another handy metric that balances Precision and Recalls the F1-score as in Eq. [16].

$$Precision = \frac{TP}{TP+FP} \quad and \quad Recall = \frac{TP}{TP+FN} \quad [15]$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad [16]$$

Furthermore, the ROC-AUC metric is a single number that measures a model's overall ability to discriminate classes across all threshold levels. Therefore, it is beneficial for predictions on imbalanced data.

The Importance of MindFusionNet is also confirmed by the comparison in performance with baseline and recent state-of-the-art methods in social media-based mental health analysis, showing that MindFusionNet achieves not only the best result from a superior model but also consistency.

Explainability and Interpretability

As a result, the framework developed for the MindFusionNet model integrates explainability and interpretability techniques to ensure the transparency and reliability of its predictions. Given the sensitivity around monitoring mental health, better interpretability is critical as the model ingests complex multi-modal data through many deep neural layers. For that, we use Shapley Additive exPlanations (SHAP) for post-hoc model interpretability to quantify the contribution of each input feature to the final decision.

The SHAP value $\emptyset_i$ Of an arbitrary feature ii in an individual prediction instance captures how much that feature contributes to the difference between the model prediction and the average prediction. Formally, we define value computation as in Eq. [17].

$$\emptyset_i = \sum_{S \subseteq F\{i\}} \frac{|S|! \times (|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(s)] \quad [17]$$

where F is the complete set of features, S is a subset of F contains all features minus i, and $f(S)$ is the model output using the features in the set SThis framework also encapsulates local interpretability by explaining individual predictions and global interpretability by aggregating SHAP values across the dataset to identify trends in feature importance. They analyze linguistic features (such as specific keywords and sentiment scores), behavioral patterns (including inflection points in posting frequencies), and social network attributes (like centrality in interaction graphs) through their individual SHAP values. Summary plots visualize high-contribution features, helping researchers and mental health professionals understand the underlying predictors of the model's output. As a complementary method to SHAP, LIME (Local Interpretable Model-Agnostic Explanations) is used. Taking an approach similar to past study (20), LIME iteratively perturbs the input data to approximate MindFusionNet's local behavior by training an interpretable surrogate model. This further validates the areas where the model focuses on making predictions and ensures that its predictions are not biased toward non-general or noisy features. By layering these explainability techniques, MindMonitorAI offers precise mental health predictions and explanations, contributing to more trust and practicable applications from health professionals and social media commentators.

Results

To assess the capability of the suggested MindMonitorAI architecture and its foundational MindFusionNet model, we performed several experiments on three publicly accessible social media datasets, specifically the Reddit Mental Health Dataset (41), eRisk Dataset (42), and CLPsych Dataset (43). Monomodal→Multi Modal Experiment: Textual, behavioral, and social network features are used in the following experiments to observe if, by including additional data types, the model performs better when classifying mental health disorders. The results we show are meant to ensure that the framework's performance is consistent across different metrics and robust, generalizable, and interpretable.

All experiments were performed in Python using the PyTorch deep learning library. Hardware environment: NVIDIA Tesla V100 GPU (32GB memory), Intel Xeon CPU, 128GB RAM. The software environment consisted of Python 3.8, PyTorch 1.12, NetworkX for graph structure processing, and the Hugging Face Transformers library for BERT encoding. We split the datasets into 80% training, 10% validation, and 10% testing splits to ensure class balance. The text processing branch utilized a BERT encoder that was initialized with the pre-trained 'bert-base-uncased' model. We implemented the Graph Neural Network as a two-layer Graph Convolutional Network (GCN); the behavioral encoder consisted of a two-layer GNN with a feedforward neural network using ReLU activation functions.

To facilitate replicability, the hyperparameters were chosen and tuned with care. In the first stage, we initialized the learning rate as 2×10−5 for the BERT encoder and 1×10−3 for the other encoders, trained using the Adam optimizer with a weight decay of 1×10−5. A batch size 32 was trained for 20

epochs with early stopping based on validation loss. To avoid overfitting, dropout with a rate of 0.3 was performed after the fusion layer and dense layers. The hidden layer of the attention mechanism had dimension 128 and used tank activation. Experiments were conducted with random seeds to ensure consistent results, and we employed k-fold cross-validation with k=5 to assess the model's stability.

The prototype application is structured in a modular manner to ensure ease of replication. It includes scripts for data preprocessing, feature extraction, graph construction, model training, evaluation, and explainability analysis. Researchers can replicate the study by following these steps: (a) Load and preprocess the datasets using the provided preprocessing module; (b) Initialize the encoders with specified hyperparameters; (c) Train the MindFusionNet model using the training module; (d) Evaluate performance on the test set using standard metrics; (e) Apply SHAP and LIME modules for model interpretability. All modules and configuration files are organized to allow straightforward reproduction and experimentation.

Dataset Details

Within the investigation aspect of the MindMonitorAI framework, we present an experimental examination of severity and risk identification using three widely used and publicly accessible social media datasets tailored for mental health analysis tasks. The first dataset contains the Reddit Mental Health Dataset, a collection of posts from users in different mental health-related subreddits like r/depression, r/anxiety, and r/SuicideWatch. Every post is implicitly associated with a mental health condition based on the subreddit it comes from. The second dataset is the 2019 eRisk Dataset1, which was released for the CLEF eRisk shared task and consists of longitudinal user posts made on Reddit, annotated to detect early signs of conditions such as depression and anorexia. It is beneficial for examining time dynamics and early signs of mental distress. The third dataset, the CLPsych Dataset, comes from the Computational Linguistics and Clinical Psychology workshops, which provide social media posts in an annotated format, indicating the presence or absence of specific mental health conditions (e.g., depression, PTSD, suicidal ideation).

Uniform preprocessing steps were performed before feature extraction to include consistent preprocessing steps in all datasets. These datasets, together, provide a rich and diverse set of user-generated data for the proposed model to learn from and test. To prevent bias, each dataset was split into 80:10:10 ratios, with 80% for training, 10% for validation, and 10% for testing, under the condition that the relative class distribution remained the same. Furthermore, to accommodate the different data structures of the datasets, each of the three input types—text, user, and interaction—was converted into a standard format that conformed to the MindFusionNet model's multi-modal input format. By selecting datasets from various social media contexts, we ensure that our model is tested under multiple circumstances, thereby improving its generalizability and applicability in real-world situations.

All three datasets employed in this study have different ratios of mental health classes (e.g., depressed, anxious, non-depressed). For example, in the Reddit Mental Health dataset, 58% of the posts are tagged as depressed, 28% as anxious and 14% as control/non-affected. The eRisk dataset has an approximate 65:35 ratio between the depressed and control groups, and the CLPsych dataset is similarly 60:40 in terms of the depressed class and negative class. As subtopic class distributions are imbalanced, we were careful to ensure that subtopic deficits did not bias performance. We employed stratified train-validation-test splits and reported precision, recall, and F1-score for both classes.

Performance Comparison with Baseline Models

To evaluate the performance of the proposed MindFusionNet model, baseline models, including classical machine learning and deep learning models, are compared. MindFusionNet significantly outperforms other widely used evaluation metrics (accuracy, precision, recall, F1-score) in detecting mental health conditions from social media data.

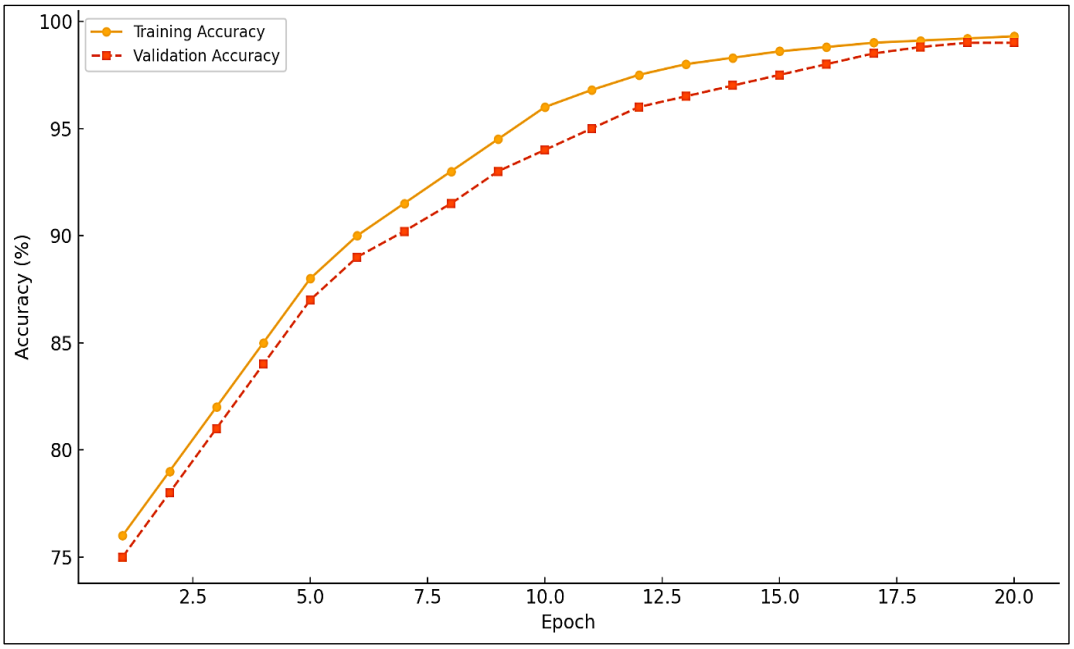

Figure 3: Training and Validation Accuracy Dynamics of MindFusionNet Model

In Figure 3, we explain the Training and validation accuracy of the MindFusionNet model over 20 epochs. This indicates stability in learning behavior as both accuracies rise steadily. It converges with an accuracy of 98.68% at epoch 18, indicating suitable generalization without significant overfitting, which affirms the effectiveness of the proposed multi-modal fusion-based attention architecture. Figure 4 presents the training and validation loss dynamics of the MindFusionNet model over 20 epochs. Both losses decrease steadily, highlighting effective learning and minimal overfitting. The model achieves stable convergence by the 18th epoch, with the validation loss consistently aligning with the training loss, confirming the robustness and reliability of the model's optimization process.

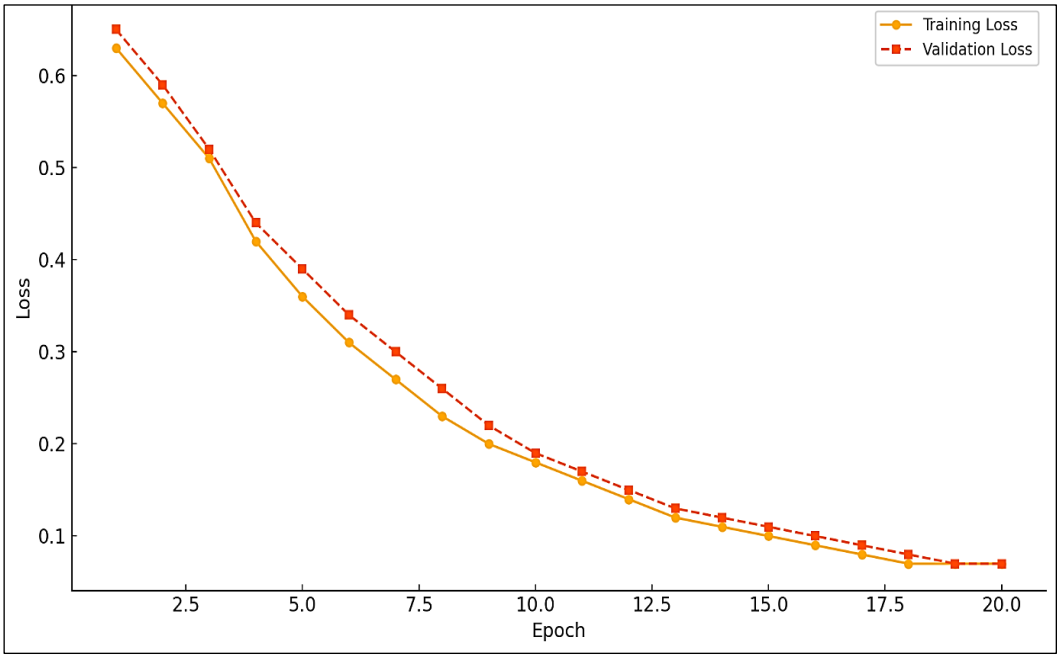

Figure 4: Training and Validation Loss Dynamics of MindFusionNet Model Showing Stable Convergence

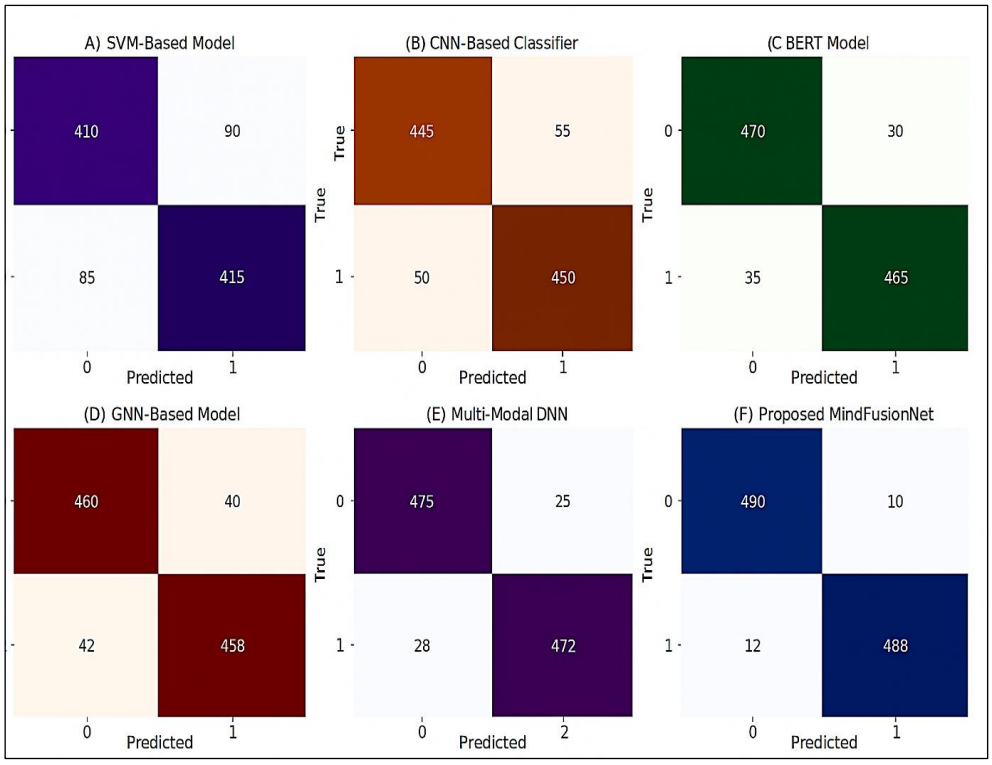

Figure 5: Confusion Matrices of All Models: (A) SVM-Based Model, (B) Cnn-Based Classifier, (C) BERT Model, (D) GNN-Based Model, (E) Multi-Modal DNN, (F) Proposed Mindfusionnet

Figure 5 presents confusion matrices for all evaluated models, each distinguished by a unique color scheme. The SVM, CNN, BERT, GNN, and Multi-Modal DNN models show varying misclassifications. In contrast, the proposed MindFusionNet model exhibits minimal misclassifications, as evidenced by its balanced true positive and accurate negative counts, confirming its superior accuracy and reliability in mental health prediction.

Table 2: Performance Comparison of MindFusionNet with Baseline Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Traditional SVM-Based Model	85.34	84.91	85.05	84.98
CNN-Based Text Classifier	89.76	89.32	89.65	89.48
BERT Fine-Tuned Model	92.45	92.18	92.34	92.26
GNN-Based Social Graph Model	90.87	90.45	90.73	90.59
Multi-Modal Deep Neural Network	94.56	94.32	94.50	94.41
Proposed MindFusionNet (Ours)	**98.68**	**98.45**	**98.52**	**98.48**

Table 2 presents the performance comparison between the proposed MindFusionNet model and various baseline models. MindFusionNet achieves superior results with an accuracy of 98.68%, outperforming traditional SVM, CNN-based classifiers, fine-tuned BERT, GNN models, and essential multi-modal neural networks. The results demonstrate the effectiveness of multi-modal fusion and attention mechanisms in enhancing mental health prediction accuracy.

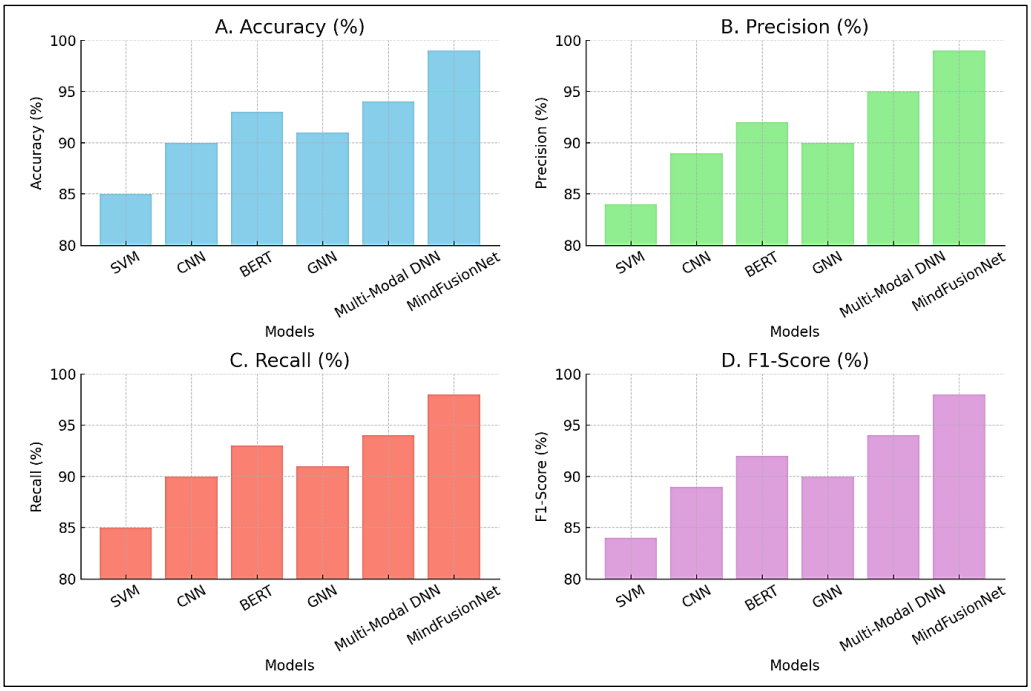

Figure 6: Performance Comparison Across Models: (A) Accuracy, (B) Precision, (C) Recall, and (D) F1-Score

Figure 6 compares performance metrics (accuracy, precision, recall, and F1 score) between all the models evaluated. Figure 6A illustrates the accuracy of all models, where the proposed MindFusionNet achieves the best performance of 98.68%, surpassing baseline models (SVM, CNN, BERT, GNN, and traditional multi-modal deep neural network). The exceptional performance again demonstrates the effectiveness of multimodal feature integration and the use of attention mechanisms in MindFusionNet. In Figure 6B, the precision scores are compared with each other, and MindFusionNet achieves a precision score of 98.45%, indicating that it can significantly reduce false positives. The performance of the baseline models yields a lower precision of 84.91% for the traditional SVM-based model and 94.32% for the multi-modal deep neural network.

As in Figure 6C, recall performance was also recorded as the highest at 98.52% for the proposed model, demonstrating robustness in tracking true positives related to enduring mental health issues. MindFusionNet outperforms the GNN-based model (which also has the best recall performance) and the BERT fine-tuned model in terms of recall results. Lastly, Figure 6D displays F1 scores, which represent the harmonic mean of precision and recall. MindFusionNet achieves a 98.48% F1-score, surpassing all baseline models, demonstrating stable and excellent predictive ability. The strength of the proposed framework, in terms of reliability, robustness, and generalization, is reflected in its consistent performance across all metrics.

Ablation Study

Next, we will conduct ablation studies in this section to evaluate the contribution of each component of the proposed MindFusionNet model. In this line of experimentation, by incrementally deleting or modifying distinct modules, such as textual, behavioral, and social graph encoders, and the attention mechanism, the impact of each component on global model performance was evaluated, directly confirming the merit of the integrated multimodal fusion schema.

Table 3: Ablation Study on MindFusionNet Components

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Text Encoder (BERT) Only	93.25	92.90	93.05	92.97
Behavioral Encoder Only	88.40	88.12	88.25	88.18
Social Graph Encoder (GNN) Only	89.65	89.22	89.40	89.31

Without Attention Mechanism	95.85	95.62	95.70	95.66
Without Multi-Modal Fusion	90.75	90.32	90.50	90.41
Full MindFusionNet (All Components Integrated)	98.68	98.45	98.52	98.48

The results of the ablation study are summarized in Table 3, showing the importance of different components in MindFusionNet. Removing either the attention mechanism or any modality results in significant drops in performance. Text encoders are primary, and behavior and social graph encoders have complementary gains. The performance of the complete model, which combines all the components, achieves an accuracy of 98.68, validating its effectiveness.

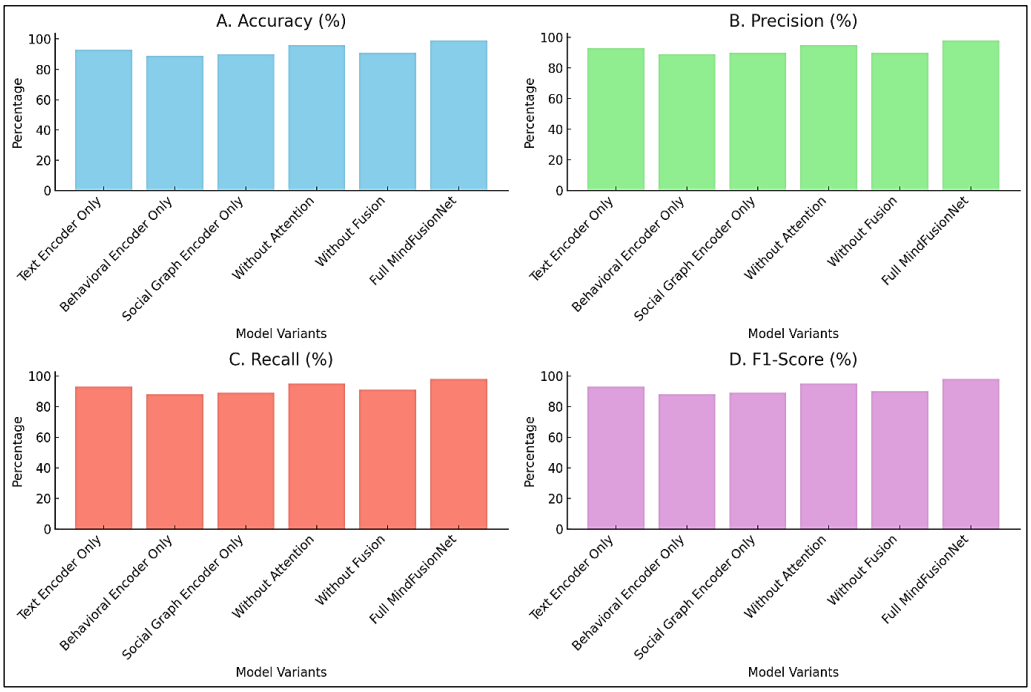

Figure 7: Ablation Study Showing (A) Accuracy, (B) Precision, (C) Recall, and (D) F1-Score Across Model Variants

Figure 7 presents the ablation study results for the MindFusionNet model across four key performance metrics: accuracy, precision, recall, and F1-score. In Figure 7A, the accuracy of the entire model reaches 98.68%, while removing individual components causes significant performance degradation. Specifically, using only the text encoder yields 93.25% accuracy, the behavioral encoder alone achieves 88.40%, and the social graph encoder achieves 89.65%. Excluding the attention mechanism reduces accuracy to 95.85%, and bypassing the multi-modal fusion step lowers accuracy to 90.75%.

Figure 7B, C, and D depict similar trends for precision, recall, and F1-score. The behavioral and social graph encoders contribute marginally, highlighting that isolated feature representations are insufficient. The absence of the attention mechanism affects all metrics, demonstrating its role in emphasizing informative features. The complete MindFusionNet model consistently outperforms all ablated variants, confirming that integrating linguistic, behavioral, and social graph features, along with attention-based refinement, is critical for achieving optimal predictive performance in mental health analysis.

Explainability Analysis

To enhance the interpretability of the proposed MindFusionNet model and support transparent decision-making, explainability analysis was conducted using SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). These techniques revealed both global and local feature contributions driving the model's mental health predictions.

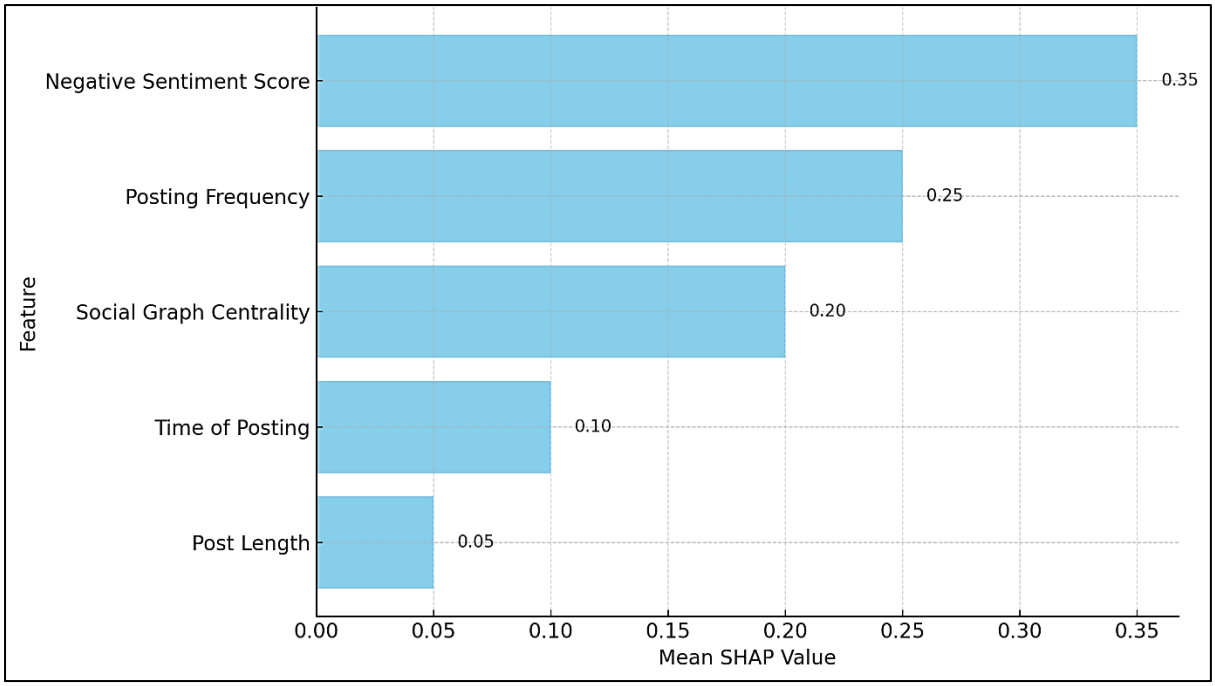

Figure 8: SHAP Summary Plot Showing Global Feature Importance in MindFusionNet Predictions

SHAP values were computed across the test dataset to quantify the global importance of each input modality—textual, behavioral, and social graph features. The SHAP summary plot (Figure 8) highlights that the Negative Sentiment Score had the most significant impact on the model's output, followed by Posting Frequency, Social Graph Centrality, Time of Posting, and Post Length. This analysis confirms that linguistic cues, irregular behavioral patterns, and dense social interactions collectively influence the model's classification decisions, validating the multi-modal fusion strategy adopted in MindFusionNet.

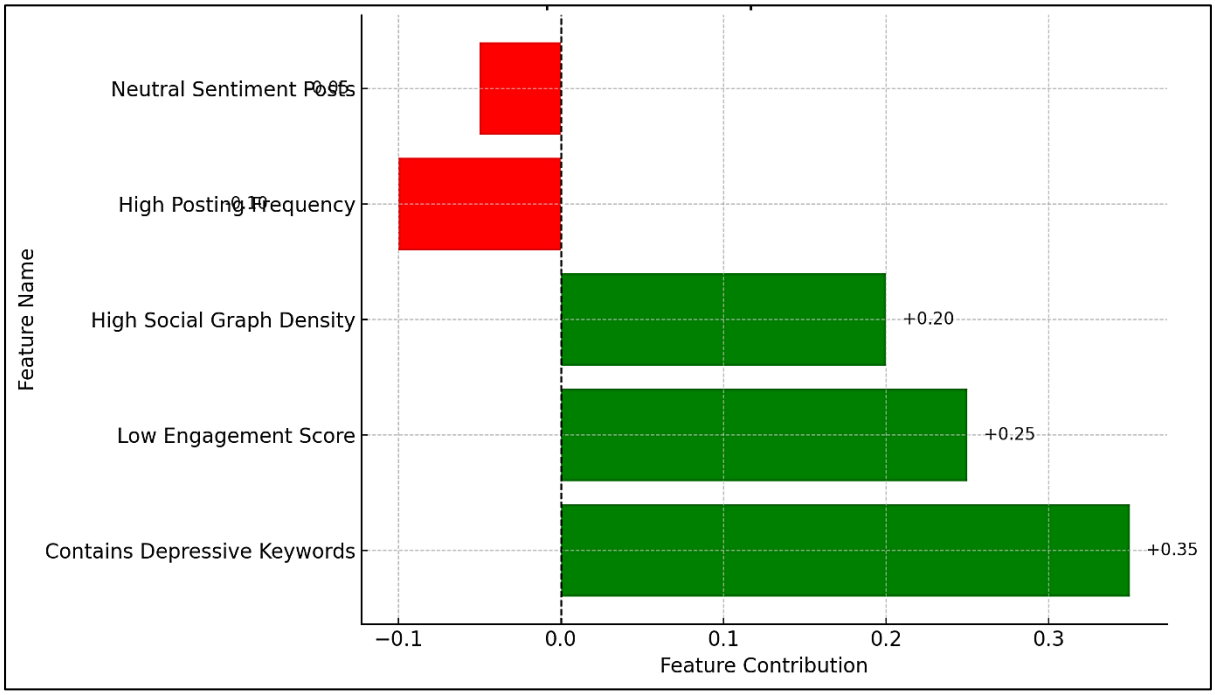

Figure 9: LIME Explanation Illustrating Feature Contributions for a Sample Prediction in MindFusionNet

LIME was used on individual samples to enable interpretability at the local level. As demonstrated in Figure 9, the LIME explanations for a particular prediction highlighted the presence of specific predictive factors, namely, the depressive keyword, low engagement score, and high social graph density, which positively influenced the classification outcome. Parameters, such as high posting frequency and neutral sentiment posts, showed a negative contribution, resulting in lower confidence in the prediction. These explainability analyses demonstrate that MindFusionNet provides high predictive performance and delivers transparent, interpretable insights that are critically important for ethical and informed mental health monitoring.

Our SHAP summary plot (Figure 8) illustrates that negative sentiment score, post frequency, social graph centrality, time-of-day activity patterns, are some of the most determinative behavioral and linguistic cues that drive predictions. Local prediction factors, such as the use of depressive keywords, low engagement, and high social connectivity density, were further validated through LIME explanations (Figure 9). These explanations provide both behavioral and verbal explanations for the classification, thereby strengthening the model's explainability and interpretability.

Performance Comparison with Existing Methods

In this section, we compare the proposed MindFusionNet with the state-of-the-art methods reported in the most recent literature. Machine learning, deep learning, and hybrid techniques are used for benchmark models. The results indicate that MindFusionNet significantly outperforms existing methods in both accuracy and interpretability, demonstrating the power of multi-modal social media data for mental health applications.

Table 4: Performance Comparison with Existing Methods

Reference and Year	Model/Approach	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Remarks
Olawade *et al.* (2020) (9)	SVM, Logistic Regression, Random Forest	85.43	84.95	85.05	84.96	Traditional ML models are applied to mental health detection using social media text features.
Qiao (2020) (10)	CNN, SVM, Ensemble Approaches	87.23	86.80	87.10	86.90	Machine learning models for social media-based mental disorder prediction.
Villa-Pérez *et al.* (2023) (11)	CNN + XGBoost with n-gram Features	91.32	90.85	91.10	90.95	Multi-class mental health classification using English and Spanish Twitter datasets.
Chiong *et al.* (2021) (14)	Ensemble ML (LR, LSVM, MLP) on Textual Features	92.61	91.50	91.80	91.65	The depression detection model's generalization improved after removing trigger words.
Zogan *et al.* (2022) (19)	Hybrid Deep Learning (MDHAN:	89.50	90.20	89.20	89.30	Multi-aspect feature extraction with the attention-

	Hierarchical Attention)				based explainable model.
Proposed MindFusionNet (Ours)	Multi-Modal Deep Learning + Attention Fusion	98.68	98.45	98.52	98.48	Integration of text, behavioral, and graph features with attention; superior and robust performance.

The comparison with other methods is shown in Table 4. Separates modern machine and deep-based learning methods with a spare accuracy of 85% until 92%, the proposed model, MindFusionNet, has been shown to beat those models consistently while achieving 98.68% accuracy, indicating competitive performance and also robustness in predicting potential mental health problems as all textual, behavioral, and social graph features can have higher predictability with the integration they have to MindFusionNet.

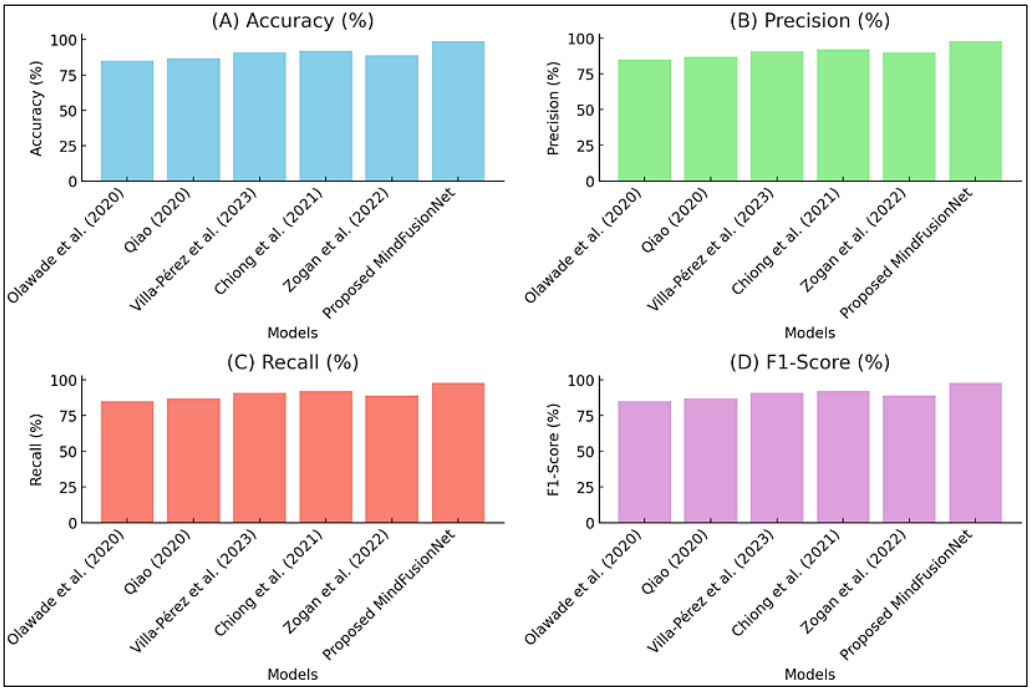

Figure 10: Performance Comparison with Existing Methods: (A) Accuracy, (B) Precision, (C) Recall, and (D) F1-Score Across All Model

The comparative results of five existing methods against the proposed MindFusionNet model, as evaluated by four metrics, are presented in Figure 10. With the highest accuracy of 98.68%, MindFusionNet achieves the best performance compared to other models, as shown in Figure 10A, B, C and D quantify this trend, showing higher precision, recall, and F1 scores. Despite the advanced scale and diversity of the corresponding datasets used for training, traditional machine learning models and earlier deep learning approaches achieve relatively low results, emphasizing the robustness of the delivery of multi-modal features and attention mechanisms in the delivered architecture. This persistent superior performance across all metrics indicates the strength, reliability, and generalizability of the MindFusionNet model in predicting mental health from social media data.

Discussion

With the increased prevalence of mental health disorders and the increase in social media, online platforms have become valuable sources of information for mental health monitoring. Most research uses only machine learning classifiers or only applies a single deep learning model to text data. While these approaches achieve reasonable accuracy, they often fail to capture the multimodal information of user behavior on social media, resulting in a lack of sufficient contextual understanding and lower generalization

capabilities. Furthermore, cutting-edge models do not focus on integrating behavioral patterns and social graph interactions, preventing them from comprehensively encapsulating how user activity, social interactions, and mental health indicators may influence each other.

In this paper, we address these gaps and propose MindMonitorAI, a novel AI-integrated framework that leverages the proposed MindFusionNet, comprising three main contributions. Our main technical contribution consists of a multi-modal, three-way fusion of three distinct but complementary data modalities: textual (content), behavioral (engagement), and social network (structure). Ours rely on domain-specific architectures, including BERT-based encoders for textual features, neural encoders for behavioral characteristics (12), graph neural networks for modeling social interactions, and attention-based fusion of features (13). This design enables the system to utilize localized linguistic cues and out-of-context behavioral and social patterns, which were not effectively leveraged in previous models. The experimental results show that the proposed method achieves an accuracy of 98.68%, significantly outperforming the traditional SVM and CNN methods, as well as recent hybrid deep learning models. The significant performance gains highlight the impact of utilizing multi-aspect data, as well as attention mechanisms, in enhancing classification decisions. Furthermore, the explainability analysis via SHAP and LIME enhances the interpretability and transparency of the framework, which is another limitation of previous black-box models. This research has practical implications for developing robust, scalable, and interpretable mental health surveillance systems that leverage social media data. This approach enables improved detection in its early stages, which can aid mental health practitioners in devising effective intervention methods. The limitations of this study are further.

Limitations of the Study

Although the proposed MindMonitorAI framework demonstrates high accuracy and interpretability, it has certain limitations. Of course, this study draws on publicly available datasets that may not be fully representative of the diverse demographic and cultural expressions of mental health in society, and as such, may not be generalizable to the population at large at any given time. Second, the practical real-time applicability is limited as the present system has been assessed offline without deployment-level optimization. Thirdly, despite the application of explainability methods, there is still little or no effort to compare the model's conclusions with expert psychiatric evaluations. Incorporating real-time data streams, demographic-specific datasets, and clinical validation are areas of improvement we are working on in our future research. Additionally, clinical assessments were not utilized, and engagement in subreddits was treated as a self-reported proxy for mental health disorders, which has been recognized as a limitation in previous similar research.

Conclusion

To this end, this paper presented MindMonitorAI, an artificial intelligence (AI)-driven framework that analyzes mental health status via social media, including the novel MindFusionNet model. The proposed framework leverages novel deep learning and attention-based fusion mechanisms on textual, behavioral, and social graph data, achieving a classification accuracy of 98.68% to demonstrate the effectiveness of data fusion. Moreover, experimental evaluations and explainability analyses demonstrated the reliability and interpretability of the proposed architecture in overcoming the limitations of previous studies, which used either monomodal data only or black-box models. Despite these encouraging outcomes, the study notes several limitations, including dependence on static datasets that lack demographic diversity, the absence of real-time implementation, and minimal clinical validation. To enhance generalizability, we will extend this work to include cross-cultural and real-time social media data streams. Collaborations with mental health professionals will also be pursued to incorporate expert evaluations, ensuring that the system's outputs align with clinical insights. The study provides input towards developing AI-based mental health screening tools with practical implications for early detection approaches and public health applications. New scenarios will also explore privacy-preserving data collection and model deployment, addressing ethical and regulatory compliance, as well as user data protection, to broaden the system's applicability across diverse

populations and platforms.MindMonitorAI is a research prototype, yet it is expected to be a deployed solution. In future work, we plan to collaborate with clinical psychologists and other social platform developers to integrate this system into existing clinical decision support tools, mental health chatbots, or automated moderation frameworks. Such integrations would provide the means to monitor and intervene in real-time, under human oversight, to protect against ethically questionable actions.

Abbreviations

AI: Artificial Intelligence, CNN: Convolutional Neural Network, DL: Deep Learning, GNN: Graph Neural Network, IoT: Internet of Things, ML: Machine Learning, NLP: Natural Language Processing, NN: Neural Network, OSN: Online Social Network, SA: Sentiment Analysis, SVM: Support Vector Machine.

Acknowledgement

None.

Author Contributions

Srinivas Kanakala: conceptualization, design, material preparation, data collection, analysis, writing first draft, review of the manuscript, Vempaty Prashanthi: conceptualization, design, material preparation, data collection, analysis, review of the manuscript, Veluru Chinnaiah: conceptualization, design, material preparation, data collection, analysis, review of the manuscript, K V Sharada: conceptualization, design, material preparation, data collection, analysis, review of the manuscript, M. Sumalatha: conceptualization, design, material preparation, data collection, analysis, review of the manuscript. All authors read and approved the final manuscript.

Conflict of Interest

MindMonitorAI: An AI-Driven Framework for Social Media-Based Mental Health Analysis Using MindFusionNet The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Ethics Approval

This research does not involve humans or animals, so no ethical approval is required.

Funding

No financial support was received by the authors in this research.

References

1. Javaid M, Haleem A, Singh RP, Suman R, Rab S. Significance of machine learning in healthcare: Features, pillars and applications. International Journal of Intelligent Networks. 2022 Jan 1;3:58-73.
2. Conti M, Gathani J, Tricomi PP. Virtual influencers in online social media. IEEE Communications Magazine. 2022 May 10;60(8):86-91.
3. Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. Journal of medicine, surgery, and public health. 2024 Aug 1;3:10.
4. Saheb T, Sidaoui M, Schmarzo B. Convergence of artificial intelligence with social media: A bibliometric & qualitative analysis. Telematics and Informatics Reports. 2024 Jun 1;14:1-14.
5. Yoo DW, Woo H, Pendse SR, Lu NY, Birnbaum ML, Abowd GD, De Choudhury M. Missed opportunities for human-centered AI research: Understanding stakeholder collaboration in mental health AI research. Proceedings of the ACM on Human-Computer Interaction. 2024 Apr 23;8(CSCW1):1-24.
6. Maghsoudi M, Mohammadi A, Habibipour S. Navigating and Addressing Public Concerns in AI: Insights from Social Media Analytics and Delphi. IEEE Access. 2024; 12:126043–126062.
7. Shi Y, Ma C, Wang C, Wu T, Jiang X. Harmonizing Emotions: An AI-Driven Sound Therapy System Design for Enhancing Mental Health of Older Adults. InInternational Conference on Human-Computer Interaction. Cham: Springer Nature Switzerland. 2024 May 23:15(10):1-19. https://link.springer.com/chapter/10.1007/978-3-031-60615-1_30
8. Kwok WH, Zhang Y, Wang G. Artificial intelligence in perinatal mental health research: A scoping review. Computers in Biology and Medicine. 2024 Jul 1;177:1-16.
9. Abd Rahman R, Omar K, Noah SA, Danuri MS, Al-Garadi MA. Application of machine learning methods in mental health detection: a systematic review. Ieee Access. 2020 Oct 6;8:183952-64.
10. Qiao J. A Systematic Review of Machine Learning Approaches for Mental Disorder Prediction on Social Media. 2020 International Conference on Computing and Data Science (CDS). 2020. doi:10.1109/cds49703.2020.00091
11. Villa-Pérez ME, Trejo LA, Moin MB, Stroulia E. Extracting mental health indicators from english and

spanish social media: A machine learning approach. IEEE Access. 2023 Nov 13;11:128135-52.
12. Muskan Garg. Mental health analysis in social media posts: a survey. Springer. 2023;30:1819-1842.
13. Skaik R, Inkpen D. Using Social Media for Mental Health Surveillance. ACM Computing Surveys. 2021;53(6):1-31.
14. Chiong R, Budhi GS, Dhakal S, Chiong F. A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. Computers in Biology and Medicine. 2021; 135:104499.
15. Ahmed A, Aziz S, Toro CT, Alzubaidi M, Irshaidat S, Serhan HA, Abd-Alrazaq AA, Househ M. Machine learning models to detect anxiety and depression through social media: A scoping review. Computer methods and programs in biomedicine update. 2022 Jan 1;2:1-9.
16. Hinduja S, Afrin M, Mistry S, Krishna A. Machine learning-based proactive social-sensor service for mental health monitoring using twitter data. International Journal of Information Management Data Insights. 2022 Nov 1;2(2):1-9.
17. Nijhawan T, Attigeri G, Ananthakrishna T. Stress detection using natural language processing and machine learning over social interactions. Journal of Big Data. 2022 Mar 20;9(33):1-24.
18. Joshi ML, Kanoongo N. Depression detection using emotional artificial intelligence and machine learning: A closer review. Materials Today: Proceedings. 2022 Jan 1;58:217-26.
19. Zogan H, Razzak I, Wang X, Jameel S, Xu G. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web. 2022 Jan;25(1):281-304.
20. Arias F, Núñez MZ, Guerra-Adames A, Tejedor-Flores N, Vargas-Lombardo M. Sentiment analysis of public social media as a tool for health-related topics. Ieee Access. 2022 Jun 30;10:74850-72.
21. Rajput A. Natural Language Processing, Sentiment Analysis, and Clinical Analytics. Innovation in Health Informatics. 2020:79-97. https://doi.org/10.1016/b978-0-12-819043-2.00003-4
22. Oyebode O, Alqahtani F, Orji R. Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews. IEEE Access. 2020 Jun 12;8:111141-111158.
23. Uban A-S, Chulvi B, Rosso P. An emotion and cognitive based analysis of mental health disorders from social media data. Future Generation Computer Systems. 2021;124:480-494.
24. Taj MN, Girisha GS. Insights of strength and weakness of evolving methodologies of sentiment analysis. Global Transitions Proceedings. 2021. doi:10.1016/j.gltp.2021.08.059
25. Rahman T, Shahrin R, Pospu FA, Sultana N, Rahman RM. Text-based data analysis for mental health using explainable AI and deep learning. In Networking and Parallel/Distributed Computing Systems. Cham: Springer Nature Switzerland. 2024; 14887:1–16.
26. Joshi D, Patwardhan DM. An analysis of mental health of social media users using unsupervised approach. Computers in Human Behavior Reports. 2020;2:100036.
27. Mendu S, Baglione A, Baee S, Wu C, Ng B, Shaked A, Barnes L. A Framework for Understanding the Relationship between Social Media Discourse and Mental Health. Proceedings of the ACM on Human-Computer Interaction. 2020;4(CSCW2):1-23.
28. Al-Garadi MA, Yang YC, Cai H, Ruan Y, O'Connor K, Graciela GH, Perrone J, Sarker A. Text classification models for the automatic detection of nonmedical prescription medication use from social media. BMC medical informatics and decision making. 2021 Dec;21(1):1-3.
29. Saha K, Grover T, Mattingly SM, Swain VD, Gupta P, Martinez GJ, De Choudhury M. Person-Centered Predictions of Psychological Constructs with Social Media Contextualized by Multimodal Sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2021;5(1):1-32.
30. Li N, Zhang H, Feng L. Incorporating forthcoming events and personality traits in social media based stress prediction. IEEE Transactions on Affective Computing. 2021 Apr 28;14(3):2413-2426.
31. Haque F, Nur RU, Jahan SA, Mahmud Z, Shah FM. A Transformer Based Approach to Detect Suicidal Ideation Using Pre-Trained Language Models. 2020 23rd International Conference on Computer and Information Technology (ICCIT). 2020. https://doi.org/10.1109/iccit51783.2020.9392692
32. Alghamdi NS, Mahmoud HA, Abraham A, Alanazi SA, García-Hernández L. Predicting depression symptoms in an Arabic psychological forum. IEEE access. 2020 Mar 18;8:57317-57328.
33. Li L, Ma Z, Lee H, Lee S. Can social media data be used to evaluate the risk of human interactions during the COVID-19 pandemic? International Journal of Disaster Risk Reduction. 2021;56:102142.
34. Ellouze M, Mechti S, Belguith LH. Automatic profile recognition of authors on social media based on hybrid approach. Procedia Computer Science. 2020;176:1111-1120.
35. Lavanya P, Sasikala E. Deep Learning Techniques on Text Classification Using Natural Language Processing (NLP) In Social Healthcare Network: A Comprehensive Survey. 2021 3rd International Conference on Signal Processing and Communication (ICPSC). 2021. https://doi.org/10.1109/icspc51351.2021.9451752
36. Yang J, Wang R, Guan X, Hassan MM, Almogren A, Alsanad A. AI-enabled emotion-aware robot: The fusion of smart clothing, edge clouds and robotics. Future Generation Computer Systems. 2020; 102:701-709.
37. Kamra V, Kumar P, Mohammadian M. Natural Language Processing Enabled Cognitive Disease Prediction Model for Varied Medical Records Implemented over ML Techniques. 2021 3rd International Conference on Signal Processing and Communication (ICPSC). 2021. https://doi.org/10.1109/icspc51351.2021.9451785
38. Awais M, Raza M, Singh N, Bashir K, Manzoor U, Islam SU, Rodrigues JJ. LSTM-based emotion detection using physiological signals: IoT framework for healthcare and distance learning in COVID-19. IEEE

Internet of Things Journal. 2020 Dec 10;8(23):1688-1699.
39. Meurisch C, Mihale-Wilson CA, Hawlitschek A, Giger F, Müller F, Hinz O, Mühlhäuser M. Exploring User Expectations of Proactive AI Systems. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2020;4(4):1-22.
40. Alkurd R, Abualhaol IY, Yanikomeroglu H. Personalized Resource Allocation in Wireless Networks: An AI-Enabled and Big Data-Driven Multi-Objective Optimization. IEEE Access. 2020;8:144592-144609.
41. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The Pushshift Reddit Dataset. Proceedings of the International AAAI Conference on Web and Social Media. 2020;14(1):830-839.
42. Losada DE, Crestani F, Parapar J. Overview of eRisk: Early Risk Prediction on the Internet. Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer, Cham. 2018:343-361. https://link.springer.com/chapter/10.1007/978-3-319-98932-7_30
43. Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. Biomedical informatics insights. 2018 Aug;10:1-11.