IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 October 2024, accepted 4 December 2024, date of publication 17 December 2024, date of current version 22 January 2025.

Digital Object Identifier 10.1109/ACCESS.2024.3519308

RESEARCH ARTICLE

A Hybrid Transformer Architecture for Multiclass Mental Illness Prediction Using Social Media Text

ADNAN KARAMAT[1], MUHAMMAD IMRAN[1], MUHAMMAD USMAN YASEEN[1], RASOOL BUKHSH[1], SHERAZ ASLAM[2,3], AND NOUMAN ASHRAF[4], (Member, IEEE)

[1]Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad 45550, Pakistan
[2]Department of Electrical Engineering, Computer Engineering and Informatics, Cyprus University of Technology, 3036 Limassol, Cyprus
[3]Department of Computer Science, CTL Eurocollege, 3077 Limassol, Cyprus
[4]School of Electrical and Electronic Engineering, Technological University Dublin, Dublin, D02 HW71 Ireland

Corresponding authors: Nouman Ashraf (nouman.ashraf@tudublin.ie), Muhammad Imran (mimran@comsats.edu.pk), and Sheraz Aslam (sheraz.aslam@cut.ac.cy)

ABSTRACT Mental illness prediction through text involves employing natural language processing (NLP) techniques and deep learning algorithms to analyze textual data for the identification of mental disorders. Therefore, machine learning and deep learning algorithms have been utilized in the existing literature for the detection of mental illness. However, current systems exhibit suboptimal performance primarily due to their reliance on traditional embedding techniques and generic language models to generate text embeddings. To address this limitation, there is a requirement for domain-specific pretrained language models that comprehensively understand the context found in posts of person with a psychiatric disability patients. Posts from individuals with mental illness often contain metaphorical expressions, posing a challenge for existing models in understanding such figurative language. In this study, we propose a hybrid transformer architecture, comprising MentalBERT and MelBERT pretrained language models, cascaded with CNN models to generate and concatenate deep features. MentalBERT is pretrained on an extensive corpus of text data specifically related to the mental health domain, while MelBERT is trained on a large corpus of metaphorical data for improved understanding of metaphorical expressions. The results reveal outstanding performance of the proposed architecture with an overall accuracy of 92% and an F1-score of 92%, surpassing state-of-the-art models in comparison. This study underscores the necessity for further research in this field and illustrates the potential of advanced technologies to address mental health issues in contemporary society.

INDEX TERMS Convolutional neural network, deep learning, mental illness, MentalBERT, melBERT, social media, transformer.

I. INTRODUCTION

Mental health disorders, also known as mental illnesses, include a wide range of conditions that affect a person's life, mood, conduct and thought processes. Depression, anxiety, stress, eating disorders, and addictive behaviors are a few instances of mental illnesses [1], [2], [3]. The World Health Organization (WHO) reports that globally, one in eight people is dealing with mental health issues, with anxiety and depression being the most prevalent mental health diseases [4]. This number has significantly increased as a result of the COVID-19 pandemic that began in 2020 [4]. There are several approaches for diagnosing and treating mental health issues. Traditionally, mental illness is diagnosed by medical health professionals, like doctors, therapists, and counselors [5], [6], [7].

Social media has changed societal interaction over the last ten years. It provides an open forum for people to express their views, share personal experiences, and frequently seek support. People successfully communicate their daily activities, experiences, hopes, and emotions online and produce a significant amount of data beyond just sharing news and information [8]. This textual data can be used to create systems that predict an individual's mental well-being.

The associate editor coordinating the review of this manuscript and approving it for publication was Barbara Guidi.

Traditional studies have demonstrated a correlation between a user's social media behavior and their mental health status [9], [10] Many machine learning (ML) and deep learning (DL) models have been proposed to classify mental health using text data from social media networks [11], [12]. Building on these advancements, researchers are employing machine learning models and investigating user-generated content on social media to examine individuals' emotional states or mental illnesses, such as schizophrenia, depression, or anxiety [13], [14], [15]. For example, the study in [16] analyzed Facebook posts of users for depression detection, utilizing the linguistic inquiry and word count (LIWC) tool to assess linguistic and affective aspects of textual data. Additionally, ML and probabilistic models such as decision trees, random forest, XGBoost, and naive Bayes are used to identify conditions related to depression, anxiety and other mental illness types [17].

Deep learning (DL) has emerged as a powerful approach for detecting and classifying mental diseases by automatically extracting complicated patterns from large amounts of unstructured data. The adoption of DL for mental illness prediction has increased with the advent of BERT-based embedding techniques and sequential deep learning models, such as recurrent neural networks (RNN), and long short-term memory (LSTM) networks. For instance, [18] explored the use of RNN, LSTM, and their variants for mental illness prediction tasks. Similarly, [19], [20] applied transfer learning using BERT and its variants to classify the mental condition of social media users. In addition, [21], [22] explored the use of text data from patient interviews and medical history notes, applying DL techniques in psychiatry. Moreover, researchers have utilized DL methods to extract meaning from unstructured text, and emoticons, in an attempt to predict mental health conditions [23], [24], [25], [26].

Although both machine learning and deep learning models have demonstrated potential across various NLP applications, specific challenges arise when applying them to mental health classification [27]. The traditional machine learning models and embedding techniques struggle to grasp the intricate contextual information present in textual data. The nuances of language unique to expressions related to mental health often surpass the capabilities of these models. The utilization of advanced architectures such as BERT, RNNs, and BiLSTMs has demonstrated enhancements, but they still fall short of fully capturing the complex interdependencies among the diverse linguistic expressions in mental health discussions. These models often have difficulty capturing subtle contextual details, which limits their ability to provide a comprehensive understanding of the complexities inherent in mental health data. This paper introduces an innovative hybridization of domain-specific transformer models cascaded with CNN models, aimed at enhancing the accuracy of mental illness classification for patients.

In this study, we create an embedding vector with the help of two domain-specific, pretrained transformer models: MentalBERT and MelBERT. This hybrid approach is employed to capture the unique linguistic features present in text related to mental health. MentalBERT is pretrained on a large corpus of data about mental health-related information. It generates an embedding vector that is capable of capturing unique linguistic patterns and semantic subtleties pertaining to mental health. These specialized embeddings, in distinction to generic embeddings produced by models like BERT or DistilBERT help MentalBERT better understand peculiar mental health conditions. MelBERT, on the other hand, is pretrained with metaphorical data. Texts on mental health frequently use metaphorical language, describing experiences and symptoms with figurative terminologies. Since these metaphorical expressions require knowledge beyond the literal meaning of words, it can be difficult for generic models to interpret them correctly. MelBERT can produce embeddings that manage to capture the underlying meanings of these metaphorical expressions. The generated embedding vectors from MentalBERT and MelBERT are fed into separate CNN models. Each CNN model comprises three convolutional layers with pooling layers. These layers use different kernel sizes to extract distinct local features from the sentence embeddings. After extracting the main features from these layers, a concatenation layer merges the outputs from both CNN models to make the final vector. The concatenated vector is then passed through the fully connected layer to complete the network architecture. The key contributions of this work are listed below:

- A novel hybrid architecture combining MentalBERT and MelBERT language models is proposed to incorporate multiple aspects of language understanding for predicting type of mental illness. This model generates text embeddings by leveraging the capabilities of two pretrained transformer models facilitating comprehensive context learning across multiple dimensions.
- A cascaded CNN architecture is designed to extract and concatenate features from text embeddings generated by the MentalBERT and MelBERT models. This design comprises three convolutional and pooling layers, intending to extract deep features from the text embeddings.
- We compared the performance of our proposed model to state-of-the-art benchmark models from the literature and our model demonstrated superior results. The experiments employed a downsampled and balanced dataset which facilitated efficient model learning in a compute-effective manner.
- An ablation study is conducted to substantiate the performance of the proposed model. This involves testing the fine-tuning of various BERT variants, as well as deep learning models from the NLP domain such as LSTM and BiLSTM, for reasoning.

The remainder of the paper is organized as follows: Section II delves into the existing literature, while Section III introduces the proposed architecture. Experimental settings are outlined

in Section IV, and Section V elaborates on the results achieved through the proposed model. Section VI provides an in-depth discussion of the overall findings. Finally, Section VII highlights the limitations of this work and Section VIII draws conclusions from the work.

II. LITERATURE REVIEW

Many people around the world experience mental illnesses due to a range of circumstances, such as traumatic life events, societal pressures, and a lack of fulfillment in life. As a result, extensive research has been conducted on the process of its diagnosis and treatment. The development of computing technologies has, in this sense, added to these efforts in a variety of ways especially with the use of artificial intelligence, machine learning and deep learning [28]. There is considerable variation in the symptoms, intensity, and impact of mental health issues on day-to-day functioning. Some common types of mental illnesses targeted in this work are described below.

A. MENTAL ILLNESS TYPES

Fig. 1 provides sample instances of Reddit posts exemplifying mental illness types along with their labels.

1) DEPRESSION

Depression is a mental illness caused by the consistent feeling of sadness, insomnia, and a loss of interest in daily activities [29]. Depression diagnosis using social media text is the most researched area among all mental illness types [21].

2) ANXIETY

Anxiety causes a worry or fear emotion and is a symptom of many other illnesses, including panic disorder, phobias, and social anxiety disorder (also known as social phobia) [30]. There are a range of models which utilize social media data for diagnosing anxiety [30].

3) BORDERLINE PERSONALITY DISORDER (BPD)

BPD is linked to a mental health condition that defines unstable relationship patterns, strong emotional reactions and low life satisfaction [31]. The text-based analysis of BPD symptoms is widely researched in the literature [32], [33].

4) POST-TRAUMATIC STRESS DISORDER (PTSD)

PTSD symptoms include flashbacks, nightmares, intense anxiety, and persistent, uncontrollable thoughts brought on by horrific events that a person either experienced or observed [34]. There exist many studies which base their work on social media data for diagnosing PTSD patients [35].

B. MACHINE LEARNING MODELS

Traditional techniques use machine learning classifiers to categorize mental health conditions and choose optimal feature combinations to enhance efficiency. In [21], the authors explained the relationship between psychological traits and digital records of online behavior on social platforms such as Facebook, Instagram, Twitter, and others. They proposed that psychological characteristics could be predicted based on the language used on social media sites and the web pages that users liked. To forecast depression, the authors in [36] gathered the data from Twitter, distinguishing between positive and negative depression-related content. They used this data to train a machine learning model, specifically logistic regression.

The work in [37] reviews different machine learning models for predicting mental illness using social media data. It focuses on identifying individuals with mental health issues on platforms like Twitter and online forums through automated techniques such as pattern recognition and activity analysis. The results indicate that widespread, passive social media monitoring may help in identifying individuals who are at risk of depressive disorder. However, even with higher diagnosis rates, a large number of cases still go undiagnosed. In [38], authors proposed a machine learning framework for early detection of social network mental disorders (SNMDs). The framework presents a novel SNMD-based tensor model that incorporates multi-source learning and collects data from multiple social media networks. It addresses the challenges in identifying SNMDs and proposes a solution based on feature extraction from data of social network users. An evaluation survey involving 3,126 social network users demonstrated that the model can reliably identify potential cases of mental illness.

The work in [39] deployed five different machine learning models to predict levels of stress, anxiety, and depression based on data collected from workforce and jobless individuals using the depression, anxiety and stress scale (DASS 21) questionnaire. The results demonstrate that the random forest classifier was the most accurate model among the five models tested. The study by [40] predicted postpartum depression from social media language using machine learning techniques such as support vector machines (SVMs), multilayer perceptron neural networks, and logistic regression (LR). The analysis involved extracting features from social media text and categorizing them as postpartum depression, general, or depression. In [41], the authors used different machine learning classifiers, such as SVM, gradient boosting, neural networks, k-nearest neighbours, and logistic regression for mental illness prediction. Gradient boosting was found to have the highest accuracy through empirical evaluation, closely followed by neural networks. These results highlight the predictive power of individual classifiers as well as ensemble approaches for mental health disorders, providing important information for automated clinical diagnosis.

C. DEEP LEARNING MODELS

Deep neural networks are extensively employed in mental illness prediction because they can replicate the hierarchical structure of the human brain and provide a deeper emotional

No.	Reddit Post	Type
1 [21]	*"I just feel so trapped and I *have* to do something about it. I don't know where I'll go or what I'll do to get by. I just can't stay here any longer."*	Depression
2 [21]	*"I know this is long and I don't know if a lot of people will read this, but I really just want to help. I had 2 panic attacks over the end of February and the first day of March. I went to the doctor and had my blood work"*	Anxiety
3 [41]	*"Basically, I have given up on trying to make or maintain meaningful connections with others. It feels like every time I let someone in, they end up leaving or hurting me, confirming my worst fears of abandonment."*	BPD
4 [21]	*"This is probably going to incite a lot of disagreement, maybe even anger, but that's okay; i'm going to say it anyway. anyone else tired of being told that just talking about your problems will solve your ptsd?"*	PTSD

FIGURE 1. Sample instances of the Reddit posts exemplifying mental illness types [19], [33].

TABLE 1. Review of existing literature on the prediction of mental illnesses.

Author	Year	Model	Illness Type(s)	Dataset	Acc. (%)	F1-Sc. (%)	Pre. (%)
Tejaswini et al. [42]	2024	Glove+CNN	Depression	Reddit/Tweet	86.7	87	88
Aragón et al. [43]	2023	DisorBERT, MentalBERT, CCNN-GloVe, BoW-SVM	Anorexia, Self-harm, Depression	Reddit/mentalhealth	-	83	82
Seth et al. [44]	2023	BERT,RoBERTa,UATTA-EB	Depression, Anxiety, Bipolar, ADHD, PTSD, None	Reddit	85	86	-
Kabir et al. [45]	2023	SVM, BiLSTM, BERT, DistilBERT	Mild depression,Moderate depression,Severe depression	Tweets	78,78,74,86	-	-
Xu et al. [46]	2023	LLMs, FLAN-T5, GPT 3.5, Mental-Roberta	Mild depression, Moderate depression, Severe depression	Reddit	82, 83, 86	-	-
Santos et al. [47]	2023	CNN,LSTM,BERT	Depression, Anxiety	Tweets	84		85,83
Tavchioski et al. [48]	2023	MentalBERT, RoBERTa, BERT, BERTweet, GT	Depression	Reddit/Tweets	87	87	-
Martinez et al. [49]	2023	Random Forest, Decision Tree, KNN, Transformers	Depression	IberLEF	80	80	81
Ji et al. [50]	2022	MentalRoBERTa	Normal, Depression, Anxiety, Suicidal ideation	Reddit/Twitter/SMS-like	88, 81	89, 81	-
Ameer et al. [19]	2022	Bi-LSTM, BERT , XLNet , RoBERTa	Depression, Anxiety, Bipolar disorder, ADHD, PTSD	Reddit	83	83	-
Niu et al. [51]	2021	Hierarchical Context-Aware Graph Attention Model	Depression	DAIC-WOZ	90	92	-
Priya et al. [39]	2020	DT, RF, SVM, KNN	Anxiety, Depression, Stress	DASS 21	86	84	88
Arora et al. [52]	2019	Multinomial Naive Bayes	Depression, Anxiety	Tweets	78	-	-

semantic representation. Several researchers have developed neural network-based mental illness categorization models, including CNN models.

1) NEURAL NETWORK BASED MODELS FOR MENTAL ILLNESS CLASSIFICATION

Authors in [53] developed deep learning models employing BiLSTM with an attention layer to learn the linguistic markers of mental disorders across different language traits such as content, emotion, and style. They also conducted thorough analyses to gain a better knowledge of the numerous indications of mental diseases, examining how distinct linguistic patterns, emotional cues, and behavioral markers link with mental health concerns. In [28], the LSTM model was used to assess mental health using various text criteria, including sentiment, basic emotions, personal pronouns, absolutist terms, and negative words. The results demonstrate that sentiment, emotions and negative words serve as effective indicators of mental health. The survey article [54] explores the growing interest in using social media data for the early identification of mental diseases and their prevalence worldwide. It focuses on the relationship between emotions and mental health, thoroughly examining methods that combine emotion fusion with mental illness detection. In another research [55], the authors review ML and DL techniques for detecting mental illnesses and discuss the significant global impact of depression, anxiety, and PTSD on mental health. Furthermore, they describe a state-of-the-art automated model for recognizing these disorders, intending to improve efficiency and accuracy by applying linguistic data gathered from patient interviews. To overcome challenges such as informal language, short text lengths, and misspellings in social media content, the researchers in [56] propose a novel deep learning model for detecting depression in social media data. This model uses a term frequency-inverse document frequency integrated

modified information gain (TF-IDF-MIG) approach. Feature extraction is performed using an improved elephant herding algorithm, while classification is carried out using a hybrid model namely an attention-improved ReLU-based convolution neural network with long short-term memory (AIRCNN-LSTM).

In [57], the authors aim to create a deep learning model that can identify a user's mental health based on the information they submit, especially in Reddit forums. The model employs the TF-IDF strategy to generate word embedding and uses CNN for feature extraction. It accurately determines whether a user's post is indicative of specific mental illnesses such as autism, schizophrenia, bipolar disorder, depression, anxiety, or BPD. Moreover, the work in [58] demonstrated that conventional depression detection algorithms fail to accurately capture the crucial sentiment information from social media text. The proposed model, named multi-gated LeakyReLU CNN (MGL-CNN) addresses this problem and determines the category of mental disorder. The initial step identifies post-level emotion, while the second step evaluates the overall emotional state of the user by aggregating results obtained from the post-level analysis. However, the overall results obtained from this approach are suboptimal.

2) BERT BASED MODELS FOR MENTAL ILLNESS CLASSIFICATION

The work in [20] proposes a BERT-based approach for identifying and categorizing posts about mental illness on social media sites like Facebook and Twitter. The research employs a novel multiclass model incorporating a transformer-based architecture, specifically RoBERTa, to analyze users' emotions and psychology inside unstructured social media data. The study focuses on conditions such as depression, anxiety, bipolar disorder, ADHD, and PTSD. In [50], the study addresses the potential for social content analysis to aid in the early detection of mental disorders and suicidal ideation. It introduces two pretrained masked language models, MentalBERT and MentalRoBERTa, specifically designed for mental healthcare applications. These models are evaluated on mental disorder detection benchmarks, showing that domain-specific language representations enhance the performance of mental health detection tasks. The work in [59] employed the Longformer model to encode concatenated user posts and performed a causal study of depression and suicide risk by looking at self-reported postings. This research demonstrated the utility and efficacy of modeling the semantic representations of depressed patients using deep learning techniques. However, the model achieved a suboptimal performance of 62% F1-score on the M-CAMS dataset.

In [60], authors proposed a multiclass deep learning model to detect mental disorders such as depression and anxiety using social media text from Twitter. This approach outperforms the traditional binary classification techniques. However, it uses a GloVe pretrained model for generating text embedding, which is a static word embedding technique. Authors in [44] propose UATTA-EB (Uncertainty-Aware Test-Time Augmented Ensembling of BERTs) to address the challenges in mental health condition classification from an online platform, Reddit. This technique uses test-time data augmentation and uncertainty awareness to increase the calibration and reliability of predictions. By examining unstructured user data, the model categorizes six different mental health conditions: none, depression, anxiety, bipolar disorder, ADHD, and PTSD. With the use of deep learning, this method produces mental health assessments that are more reliable and accurate. The study in [45] develops a typology for identifying depression severity in social media texts by utilizing the clinical articulation of depression, which addresses issues in mental health research. This study presents a new dataset of 40,191 tweets labeled as 'non-depressed' or 'depressed,' with three severity levels for 'depressed' tweets: mild, moderate, and severe, mimicking DSM-5 and PHQ-9 assessment procedures. To guarantee quality, knowledgeable annotators assign a confidence score to every label. Using BERT and DistilBERT, strong baseline results and summary statistics are used to validate the quality of the dataset; limitations are discussed to help direct future research.

The background study reveals that mental illness prediction using social media text is a widely explored area in the computer science community. Initial research mainly employed static embedding techniques such as TF-IDF, bag of words (BOW), and GloVe, which ignores the context of words in a sentence. Furthermore, classification relied on ML techniques alone, which could not extract deep features from the text. Consequently, the models' performances were suboptimal and not suitable for real-life applications. However, with the introduction of contextual embedding techniques such as BERT and its variants, the performance of mental illness classification models began to improve. Additionally, the inclusion of deep feature extraction techniques helped improve the model's understanding of text in a mental illness context. The literature, as summarized in Table 1 further differentiates between binary and multiclass scenarios. While the binary classification of mental illness has improved, multiclass classification still needs focus due to the subtle difference between multiple mental illness types. This work aims to address the issue of accurate classification of multiclass mental illnesses using social media text.

III. PROPOSED METHODOLOGY

Fig. 2 illustrates the proposed methodology, comprising six primary components. Specifically designed for NLP-based mental health analysis, proposed model prioritizes capturing contextual understanding and sequential dependencies. It involves a multi-step process beginning with dataset acquisition from multiple sources to create a balanced dataset with target mental health conditions. The obtained data undergoes preprocessing, including data cleaning and lowercasing to prepare it for analysis. To capture semantic

and contextual relationships in the data, the model employs the domain-specific pretrained models, MentalBERT and MelBERT, to encode the text. The outputs of both pretrained models pass through the CNN models to extract high-level features. The extracted features are then concatenated to form a comprehensive feature vector. Finally, this combined feature vector is input into a classifier to predict model outcomes for depression, anxiety, BPD and PTSD.

A. DATASET ACQUISITION

The dataset used in this study is the consolidated version of the datasets obtained from two different sources. The source of the first dataset is a reputable research paper [57] and is obtained by sending a formal request through the provided form at the following URL: https://jinakim.github.io/dataset/20srep-mental. The source shared their Google Drive link from where we downloaded the dataset in CSV format. This dataset contains Reddit posts corresponding to mental health-related subreddits such as r/depression, r/anxiety and r/BPD. Although the dataset originally contained six classes, in this study, only depression, anxiety, and BPD were chosen for analysis due to their distinctive nature from the rest of the illnesses. These three types, together with PTSD, play a big role in understanding and classifying mental health problems—hence being the central subject in this research.

A secondary dataset was used in addition to the primary dataset to include a PTSD class complementary to the three classes chosen from the first dataset. This strategy was used to concentrate on the important classes being studied while enhancing the dataset's diversity. The second dataset was downloaded from the open-source HuggingFace platform using the following URL: https://huggingface.co/datasets/solomonk/reddit_mental_health_posts/tree/main. Combining PTSD with depression, anxiety and BPD integrates multiple mental health conditions in a single dataset for handling multiclass mental illness problems.

B. DATASET PREPROCESSING

Data preprocessing is required to make the input dataset compatible with ML and DL algorithms. It is a crucial stage in NLP since it helps in cleaning up the text by getting rid of errors and inconsistencies. These actions simplify the text, making it more adaptable for NLP algorithms to extract meaningful insights. Fig. 3 illustrates the data preprocessing steps used in this study.

The data preprocessing steps comprise data cleaning, which involves handling missing values, removing inconsistencies and applying a data deduplication procedure to remove redundant entries. First, we addressed the noise present in the data by removing HTML tags and URLs that could obstruct the meaningful aspect of the data. Special characters such as {, #, etc. do not help find the contextual information from the data and were also removed to clean and standardize the data. Further, we also handled the missing values from the data and removed rows that contained missing text data to maintain the integrity of our dataset. Duplicated records were also removed to ensure the consistency of the dataset.

Label encoding is performed to transform the categorical class labels in the dataset into numerical values for use by supervised learning models. The mental illness types of depression, anxiety, BPD and PTSD are encoded as 0, 1, 2 and 3 respectively. Furthermore, all text is converted to lowercase to ensure uniformity, preventing the model from treating words with different cases as separate entities. Collectively, these steps contribute to transforming mental illness post data into a standardized format, making it suitable for our task.

Employing a random downsampling strategy, we significantly reduced the sample size while preserving crucial data properties to optimize computing efficiency. This simplified dataset aligns well with our research objectives, allowing us to concentrate on specific aspects of mental health in a resource-efficient manner. Table 2 provides the specifics of the dataset distribution used in this study, comprising a total of 40000 samples. The balance was maintained with 10,000 samples for each class, where 8,450 were selected for training, 800 for validation, and 750 for testing per class. Therefore, a total of 33,800 samples were used for training, 3,200 for validation, and 3,000 for testing the proposed architecture.

TABLE 2. Dataset distribution in training, validation and test sets.

Class	Training	Validation	Testing	Total
Depression	8450	800	750	10000
BPD	8450	800	750	10000
PTSD	8450	800	750	10000
Anxiety	8450	800	750	10000
Total	**33800**	**3200**	**3000**	**40000**

C. ENCODING WITH PRETRAINED MODELS

This step involves selecting appropriate pretrained deep learning models, namely MentalBERT and MelBERT, chosen for their effectiveness in text classification tasks. These are BERT-based models that have undergone pretraining on extensive domain-specific text data, rendering them well-suited for embedding tasks within their respective domains. They are particularly chosen for their ability to analyze various linguistic and semantic features in text data to identify patterns and indicators of mental health conditions. MentalBERT is pretrained on mental health-related data obtained from Reddit and its corresponding subreddits [50]. It is trained to generate embeddings peculiar to the mental illness domain. The MelBERT model is pretrained on metaphorical data and is capable of understanding metaphorical expressions within the text [61]. It is specifically employed in this work to comprehend metaphorical expressions within posts of mental illness patients.

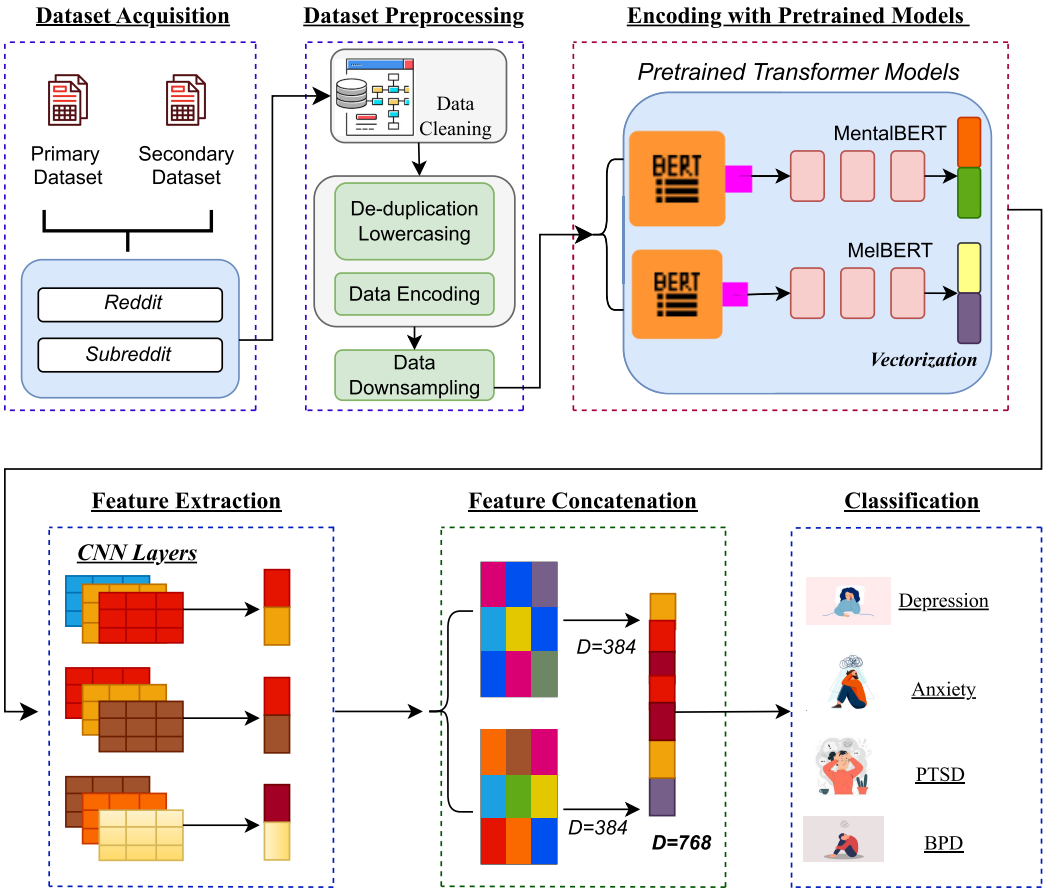


FIGURE 2. Illustration of the proposed model for mental Illness classification.

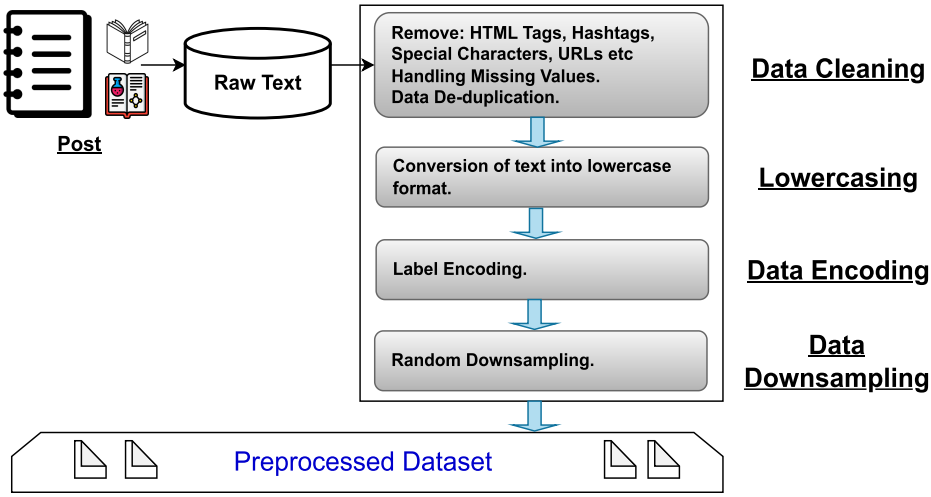


FIGURE 3. Data preprocessing pipeline used in this study.

The proposed architecture adopts a hybrid approach to generate two embedding vectors for each Reddit post input into the model. The first vector comes from the MentalBERT model, while the other one is derived from the MelBERT model. The process is illustrated in Fig. 4, which shows two branches, each taking the same post as an input and producing 768-dimensional output vectors. Both these models split the post into discrete tokens and generate embeddings using the embedding layer, resulting in a sequence of embedding vectors, each with a size of 512. Eq. 1 describes the procedure

for post, X_m, resulting in n tokens i.e. $x_{m1}, x_{m2}, \ldots, x_{mn}$.

$$X_m = [x_{m1}, x_{m2}, \ldots, x_{mn}] \tag{1}$$

Next, positional encoding is used to provide information about the position of each token in the sequence. This procedure allows the model to comprehend the order of words by incorporating positional information into input embeddings. Thus, positional encoding helps determine each word's position within the sequence, even though the embeddings themselves are static. The positional encoding vectors, $PE_{(pos,2i)}$ and $PE_{(pos,2i+1)}$, are defined in Eq. 2 and Eq. 3. Sine and cosine functions are used to capture both even and odd positions in a complementary way.

$$\mathrm{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{(2i/d_{\mathrm{model}})}}\right) \tag{2}$$

$$\mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{(2i/d_{\mathrm{model}})}}\right) \tag{3}$$

Equation 4 efficiently integrates each token's positional (PE_m) and semantic (X_m) information, allowing the Transformer model to understand the order of tokens in the sequence.

$$E_m = X_m + PE_m \tag{4}$$

The positional encoded embeddings (E_m) undergo query (Q_m), key (K_m), and value (V_m) projection. The Q_m, K_m, and V_m matrices are derived from the input embeddings through linear transformations. Eq. 5, 6, 7 defines the process of creating Q_m, K_m and V_m matrices.

$$Q_m = E_m \cdot W_Q \tag{5}$$

$$K_m = E_m \cdot W_K \tag{6}$$

$$V_m = E_m \cdot W_V \tag{7}$$

The attention function takes Q_m, K_m and V_m matrices as input and enables the BERT model to efficiently capture dependencies between tokens across an entire sequence. It serves as the basis for generating contextualized representations of input tokens, which is essential for obtaining state-of-the-art results. The attention mechanism takes the scaled dot product of Q_m and K_m matrices and calculates normalized attention weights using the softmax function. The attention weights are then used to compute the weighted sum of the V_m vector (Eq. 8).

$$\mathrm{Attention}(Q_m, K_m, V_m) = \mathrm{softmax}\left(\frac{Q_m K_m^\top}{\sqrt{d_k}}\right) V_m \tag{8}$$

The Multi-head attention combines multiple heads to focus on different parts of the input sequence, The model's capacity to identify different patterns and dependencies is enhanced as each head gains expertise in focusing on distinct facets of the input.

$$\mathrm{MultiHead}(Q_m, K_m, V_m) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W_O \tag{9}$$

Each head, head_i in Eq. 9 is computed as given in Eq. 10:

$$\mathrm{head}_i = \mathrm{Attention}(Q_m W_{i,Q}, K_m W_{i,K}, V_m W_{i,V}) \tag{10}$$

Finally, the output of multi-head attention passes through a feedforward neural network as shown in Eq. 11. The FFN refines the representations obtained from the self-attention process, thereby improving the ability to capture complex dependencies.

$$\mathrm{FFN}(x) = \mathrm{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{11}$$

The designed hybrid approach produces richer and more pertinent embeddings of person with a psychiatric disability posts by combining the strengths of the MentalBERT and MelBERT models. MentalBERT, with its emphasis on mental health-related language, provides specific expertise and contextual awareness required for effectively identifying symptoms and sentiment in patient communication. MelBERT, on the other hand, excels at metaphorical language understanding and improves the robustness of text processing due to its broad-based language modeling capabilities. By combining both models, the hybrid system can benefit from MentalBERT's specific insights and MelBERT's metaphorical comprehension, resulting in a more accurate and comprehensive analysis.

D. FEATURE EXTRACTION

The proposed architecture introduces an innovative feature extraction strategy by cascading CNN models with a hybrid BERT-based transformer architecture. This method effectively incorporates the benefits of BERT and CNNs, hence utilizing the strengths of both models. BERT efficiently extracts long-term dependencies and contextual information within sentences. In contrast, CNNs are best in feature extraction from short phrases by capturing dependencies among word combinations.

We use CNN to extract local features from the data and reduce the dimensionality of the feature vector. The core components of a CNN are the convolution layers, pooling layers and a fully connected layer. The three convolution layers apply convolution operations to the embedding vectors by sliding filters of different sizes, initially set to random values. These filters are adapted during the training to detect diverse patterns and complex features from the data. Each filter captures distinct features from the input data and creates multiple feature maps. Subsequently, the pooling layer summarizes the output from the convolutional layers and extracts significant features of the feature maps. The pooling operation intends to reduce the spatial dimensions of the feature maps while retaining the most relevant information. This reduction helps to make the computation more efficient and lowers the risk of overfitting. Furthermore, it helps convert feature maps into a fixed-sized vector, which is then used for the classification task. By utilizing filters of different sizes and pooling layers, the CNN learns to highlight the important features from the embedding vector, enhancing the overall performance of the model. Collectively,

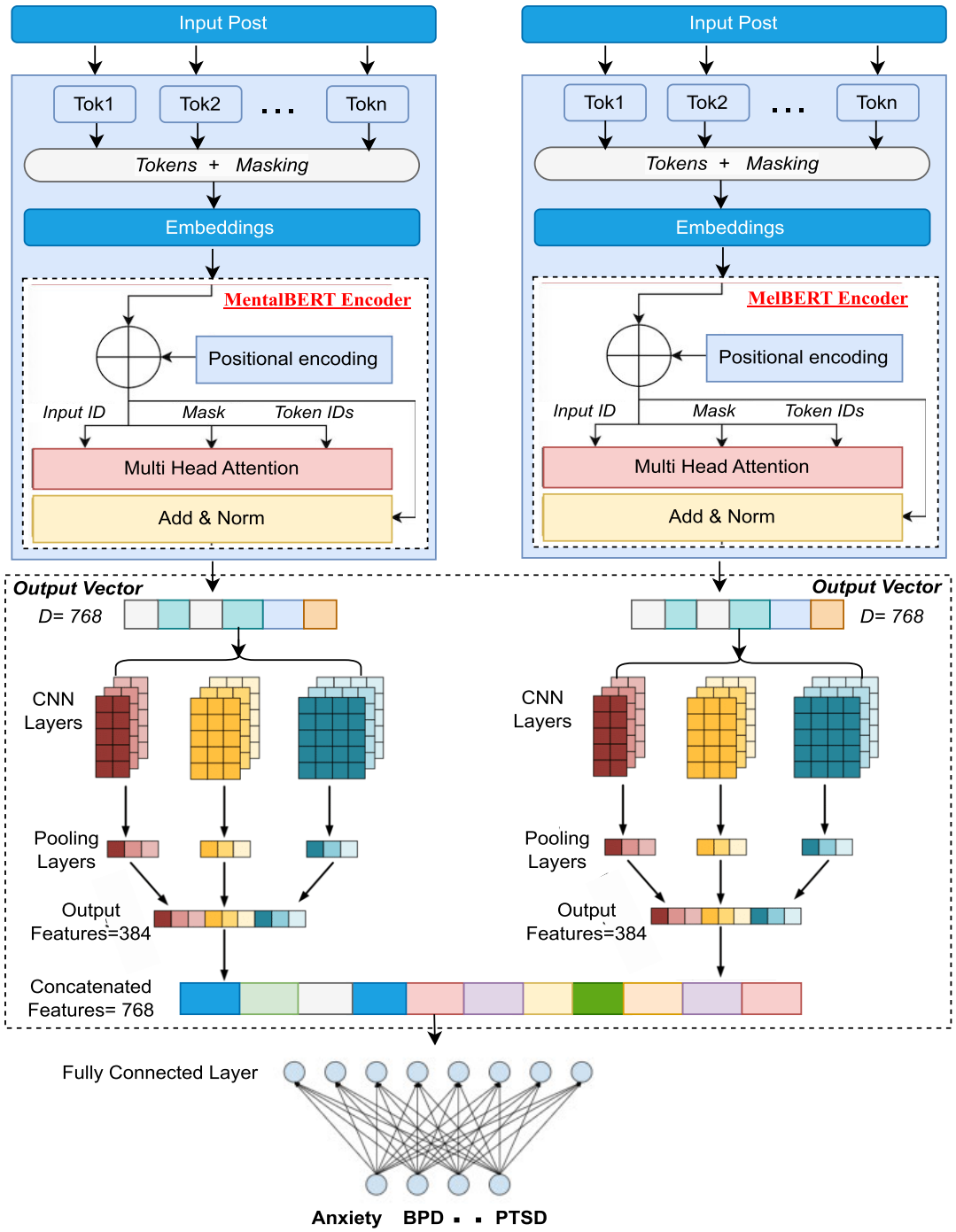


FIGURE 4. Illustration of the hybrid transformer architecture with cascaded CNNs.

the hybrid architecture and CNN model perform incredibly well, especially for handling the complex task of categorizing different types of mental illnesses.

Fig. 4 shows that the cascaded CNN architecture takes 768-dimensional embedding vectors as input from MentalBERT and MelBERT models. The CNN component in both branches comprises three convolutional and two pooling layers. The convolutional layers utilize 128 filters and kernel sizes of 2, 3, and 4, enabling effective extraction of text features. Subsequently, max-pooling layers reduce the spatial dimensions of the input by retaining the maximum value within each local region. Further enhancement is achieved through an average-pooling layer and a convolutional layer with 256 filters and a kernel size of 5, resulting in a final vector with 384 dimensions. Due to being applied in the NLP domain, these operations are implemented on the 1D

feature vector in both branches. Eq. 12 outlines generalized convolution operation for three filters.

$$C_k = W_k^{F_k} * C_{k-1} + b_k^{F_k},\ C_k \in \mathbb{R}^{L \times F_k} \text{ for } k = 1, 2, 3 \tag{12}$$

Max pooling in CNNs for textual data involves sliding a window over the input text embeddings and selecting the maximum value within each window. This process downsamples the input, reducing its dimensionality while retaining important features. This operation is shown in Eq. 13.

$$MP_k = \text{max\_pool}(C_k), \quad MP_k \in \mathbb{R}^{L/2 \times F_k} \text{ for } k = 1, 2, 3 \tag{13}$$

The reduction in dimensionality is achieved by summarizing the information in each region with average pooling layers. Equation 14 defines the process of average pooling.

$$AP_k = \text{avg\_pool}(MP_k), \quad AP_k \in \mathbb{R}^{F_k} \text{ for } k = 1, 2, 3 \tag{14}$$

where X is the input vector of 768 dimensions, W_k is the weight matrix, b_k is the bias vector and C_k represents the output feature map of kernel size k. The outcome of the average pooling operation is a 384-dimensional feature vector from both branches, each representing a more enhanced understanding of the particular mental illness.

E. FEATURES CONCATENATION

The two feature vectors, each having 384 dimensions, are concatenated to form a final 768-dimensional vector. This consolidated feature vector leverages the unique characteristics of the two models to generate an in-depth comprehension of the mental illness category. Moreover, individual feature vectors extract and refine high-level features, enabling a deeper and more nuanced understanding of the mental illness category using CNN model. Eq. 15, 16 and 17 defines the process of generating a fused feature vector, O_{fused}, by combining individual feature vectors, $O_{MentalBERT}$ and $O_{MelBERT}$ from their respective branches.

$$O_{\text{MentalBERT}} = \text{concat}(AP_1, .., AP_n),\ O \in \mathbb{R}^{F_1,..,F_n} \text{ for } n = 3 \tag{15}$$

$$O_{\text{MelBERT}} = \text{concat}(AP_1, .., AP_n),\ O \in \mathbb{R}^{F_1,..,F_n} \text{ for } n = 3 \tag{16}$$

$$O_{\text{fused}} = \text{concat}(O_{\text{MentalBERT}}, O_{\text{MelBERT}}),\ O_{\text{fused}} \in \mathbb{R}^{768} \tag{17}$$

F. MENTAL ILLNESS CLASSIFICATION

The concatenated feature vector is fed to the fully connected layer, also known as the dense layer, for classification. The softmax function is employed as the activation function for this classification process, specifically to classify into four mental illness classes: *depression*, *anxiety*, *BPD*, and *PTSD*. The final layer produces a vector of scores, where each element represents the model's confidence that the input is a member of a specific class. These scores are converted by the softmax function into a probability distribution over

Algorithm 1 Hybrid Transformer Architecture With Cascaded CNNs

Input ← Text data D, labels Y, MentalBERT model M_{Mental}, MelBERT model M_{Mel}, CNN layers L, kernel sizes K, fully connected layer parameters
Output ← Predictions, accuracy, precision, recall, F1-score
function Transformers(D, M_{Mental}, M_{Mel})
  Generate embedded vectors from MentalBERT: $V_{\text{Mental}} = M_{\text{Mental}}(D)$
  Generate embedded vectors from MelBERT: $V_{\text{Mel}} = M_{\text{Mel}}(D)$
  return V_{Mental}, V_{Mel}
end function
function CNN(V_{Mental}, V_{Mel}, L, K)
  for l in 1 to L **do**
    Apply CNN with kernel size K to V_{Mental}: $C^{(l)}_{\text{Mental}}$
    Apply CNN with kernel size K to V_{Mel}: $C^{(l)}_{\text{Mel}}$
  end for
  Concatenate C_{Mental} and C_{Mel}: $C_{\text{Concatenated}} = \text{Concat}(C_{\text{Mental}}, C_{\text{Mel}})$
  return $C_{\text{Concatenated}}$
end function
function FullyConnected($C_{\text{Concatenated}}$, fully connected layer parameters)
  Flatten $C_{\text{Concatenated}}$ into a vector
  Apply fully connected layer with specified parameters
  return Predictions
end function
Training:
Initialize model parameters
while not converged **do**
  Sample a batch of training examples (D_{batch}, Y_{batch})
  Compute embedded vectors: $V_{\text{Mental\_batch}}, V_{\text{Mel\_batch}} = \text{Transformers}(D_{\text{batch}}, M_{\text{Mental}}, M_{\text{Mel}})$
  Compute CNN features: $C_{\text{Concatenated\_batch}} = \text{CNN}(V_{\text{Mental\_batch}}, V_{\text{Mel\_batch}}, L, K)$
  Perform fully connected layer operation: $Predictions_{\text{batch}} = \text{FC}(C_{\text{Concatenated\_batch}}, \text{FC layer parameters})$
  Compute loss: $loss = \text{CrossEntropy}(Y_{\text{batch}}, Predictions_{\text{batch}})$
  Update model parameters using backpropagation and optimization algorithm
end while
Testing:
For test data D_{test}, repeat steps from Training with forward pass only
Compute accuracy, precision, recall, F1-Score for test predictions

the classes, specifying the likelihood of the input belonging to each class. The mental illness classification can be represented using Eq. 18.

$$P(y_i|O_{\text{fused}}) = \text{softmax}\left(W_f \cdot O_{\text{fused}} + b_f\right) \quad (18)$$

where $P(y_i|O_{\text{fused}})$ is the probability of the input belonging to class i. W_f is the weight matrix of the dense layer. b_f is the bias vector of the dense layer. O_{fused} is the fused feature vector obtained from the previous stages.

Algorithm 1 presents the detailed working of the proposed architecture. The approach integrates pretrained transformer models, namely MentalBERT and MelBERT, with CNNs to detect mental illnesses. The procedure involves generating text embeddings using transformers, feature extraction using CNN models, and the utilization of fully connected layers for making predictions. Model parameters are optimized through backpropagation during training, and the effectiveness is assessed on test data using standard evaluation metrics such as accuracy, precision, recall, and F1-Score. This hybrid architecture is designed to augment the model's capacity to capture complex patterns within textual information, thereby improving mental illness detection. All things considered, this methodology provides a strong base for diagnosing mental illness by analysing different linguistic and semantic aspects within textual data. This method can assist in effectively and accurately detecting mental health disorders by utilizing deep learning models and methodologies.

IV. EXPERIMENTAL SETTINGS

This work uses Python 3.10 as the primary programming language for implementation. Specifically, it uses Pandas 2.1 for data preprocessing operations, TensorFlow 2.13 and Keras 2.13 to develop deep learning models, and scikit-learn 1.2 library for performance evaluation. Furthermore, it utilizes the Kaggle computing platform, and its GPUs to train and test the models.

In this work, all experiments, including the ablation study, are conducted using a uniform dataset comprising four mental illness categories. This dataset ensures a robust and representative evaluation of our models by encompassing a wide spectrum of mental illnesses. By employing the same dataset in several experimental configurations, we preserve data distribution uniformity and allow fair comparison of the results. This method enables us to carefully evaluate the effectiveness and influence of different parts of our model architecture and ensure that any differences we see are caused by the model architectures rather than variances in the underlying dataset. Table 2 presents the dataset details used in this work.

A. HYPERPARAMETERS

The model hyperparameters along with their chosen values are presented in Table 3. The hyperparameters for MentalBERT and MelBERT are similar, with both models being $BERT_{Base}$ architectures. Each has 12 layers and 12 attention heads, 768 hidden units, and 110 million parameters. They use a decay learning rate schedule and a dropout rate of 0.1. Both models process batches of 32 with a maximum sequence length of 512 tokens.

TABLE 3. Parameter description of the MentalBERT/MelBERT with CNN and Dense network.

Model	Parameters	Value
MentalBERT/ MelBERT	No. of MentalBERT layers	12
	No. of MelBERT layers	12
	No. of attention heads	12
	Hidden units	768
	Parameters	110M
	Batch Size	32
	Max sequence length	512
	Learning Rate Schedule	Decay
	Dropout rate	0.1
CNN Model	No. of Neurons	64,128,256
	Hidden layers	03
	CNN kernel Size	03,04 ,05
	Batch size	64
	No. of Epochs	40
	Activation function	ReLU
	Padding	Same
Dense Network	Neurons	128
	Activation function	Softmax
	Optimizer	Adam
	Learning rate	0.001
	Loss function	sparse categorical crossentropy

The CNN model consists of three layers with 64, 128 and 256 neurons, utilizing kernel sizes of 3, 4, and 5 respectively. The encoded vectors from these models are sent through three hidden layers. ReLU activation and the same padding are applied throughout 40 epochs. The number of training samples in each iteration is determined by a batch size of 64, and the step size during optimization is controlled by a learning rate of 0.001. Training is conducted over 40 epochs with a learning rate schedule comprising decay and a linear warm-up, promoting steady and efficient model convergence. Finally, a dense network with 128 neurons processes the data using the softmax activation function, and the loss function is sparse categorical cross-entropy.

B. EVALUATION METRICS

One of the major metrics for assessing deep learning models, mostly in classification tasks, has to do with the confusion matrix. It provides a detailed breakdown of the model's predictions against actual labels. For a four-class mental illness classification problem, the confusion matrix is a 4×4 Table 4:

TABLE 4. Confusion matrix as a performance evaluation metric for mental illness classification.

Actual / Predicted	Depression	Anxiety	BPD	PTSD
Depression	TP_1	$FP_{1,2}$	$FP_{1,3}$	$FP_{1,4}$
Anxiety	$FN_{2,1}$	TP_2	$FP_{2,3}$	$FP_{2,4}$
BPD	$FN_{3,1}$	$FN_{3,2}$	TP_3	$FP_{3,4}$
PTSD	$FN_{4,1}$	$FN_{4,2}$	$FN_{4,3}$	TP_4

Where:

- TP_i (True Positive) is the count of correct predictions for class i.
- $FP_{i,j}$ (False Positive) is the count of instances where the model incorrectly predicted class i instead of the actual class j.
- $FN_{j,i}$ (False Negative) is the count of instances where the model incorrectly predicted class j instead of the actual class i.

Several key performance metrics can be derived from the confusion matrix as given in Eq. 19, 20, 21, 22, and 23

C. ACCURACY

The accuracy for class i is the ratio of correctly predicted true positives for all mental illness classes ($\sum_{i=1}^{4} TP_i$) to the total number of instances in the dataset. Accuracy is calculated using Eq. 19.

$$\text{Accuracy} = \frac{\sum_{i=1}^{4} TP_i}{\sum_{i=1}^{4}\sum_{j=1}^{4} M_{i,j}} \tag{19}$$

where $M_{i,j}$ is the element at the i-th row and j-th column of the confusion matrix.

D. PRECISION

Similarly, precision for class i is the ratio of TP instances of a mental illness class i to the sum of TP and the number of incorrect predictions belonging to class i. It is evaluated using Eq. 20.

$$\text{Precision}_i = \frac{TP_i}{TP_i + \sum_{j=1, j\neq i}^{4} FP_{i,j}} \tag{20}$$

E. RECALL (SENSITIVITY OR TRUE POSITIVE RATE)

Recall for mental illness class i determines the number of instances that are correctly predicted by the classifier. It is the ratio of the TP and the sum of TP and FN instances. Recall is calculated using Eq. 21.

$$\text{Recall}_i = \frac{TP_i}{TP_i + \sum_{j=1, j\neq i}^{4} FN_{j,i}} \tag{21}$$

F. F1-SCORE

F1-score for mental illness class i is the harmonic mean of their precision and recall and provides overall model performance. F1-score is calculated using Eq. 22.

$$\text{F1-Score}_i = \cdot \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{22}$$

G. FALSE POSITIVE RATE (FPR)

The FPR for mental illness class i provides the ratio of FP instances to the sum of FP and TN instances, as shown in Eq. 23.

$$\text{FPR}_i = \frac{FP_i}{FP_i + TN_i} \tag{23}$$

V. RESULTS EVALUATION AND PERFORMANCE ANALYSIS

This section presents the performance analysis of the proposed architecture and ablation experiments conducted in this study. The proposed architecture is comprehensively evaluated using training and validation curves for accuracy and loss, in addition to overall model performance. The ablation experiments involve configuring different model components to various settings for analysis. An overall comparison is conducted towards the end of the section.

A. PROPOSED HYBRID ARCHITECTURE

The proposed hybrid transformer architecture is evaluated on the acquired dataset. The results presented in Fig. 5 demonstrate that the model training was smooth, negating the presence of any overfitting issue. The accuracy curves for both training and validation demonstrate progressive enhancement with the increase in the number of epochs. The training was stopped at 40 epochs as the curves stabilized after 25 epochs. The loss curve demonstrates contrasting behavior, initially starting with high values but eventually stabilizing after 25 epochs. Overall these curves demonstrate efficient model learning with an increase in the number of epochs.

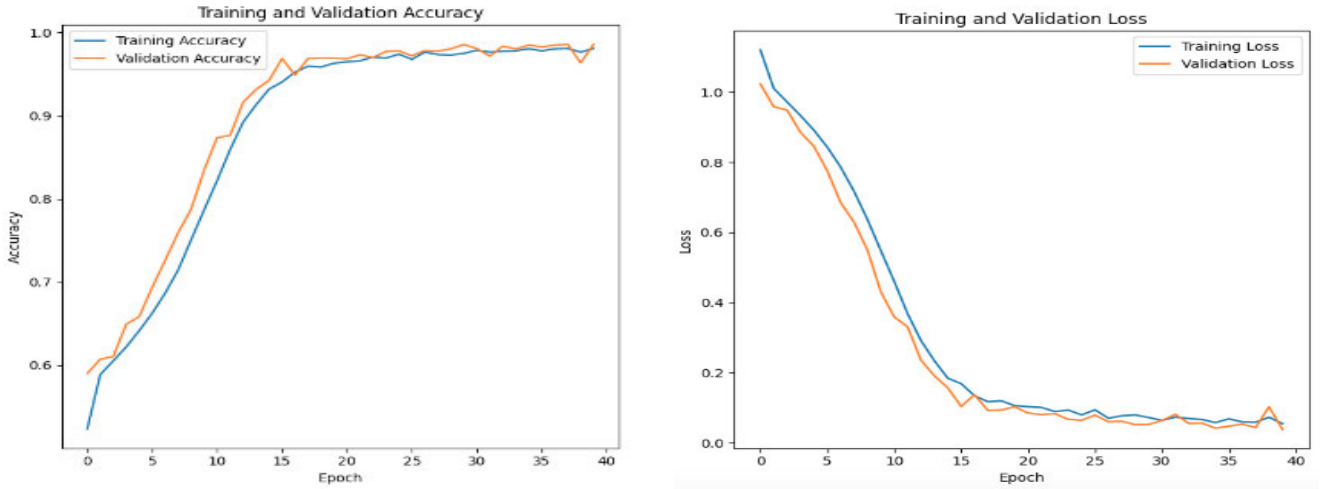
Table 5 presents a multiclass comparison of the proposed model using standard evaluation metrics. The proposed approach demonstrated superior results across all mental illness categories. The best accuracy was achieved for PTSD at 94%, while the lowest was observed for BPD at 90%. In contrast, precision was optimal for BPD and anxiety, each attaining a value of 94%. PTSD displayed the lowest precision with a value of 90%. The recall metric depicted a similar pattern to accuracy, with PTSD achieving the highest performance at 95%, while BPD exhibited the lowest at 91%.

TABLE 5. Multiclass performance evaluation of the proposed hybrid architecture using accuracy, precision, recall and F1-score metrics.

Mental Illness Type	Accuracy	Precision	Recall	F1-score
Depression	91%	93%	92%	92%
Anxiety	93%	94%	93%	94%
BPD	90%	94%	91%	92%
PTSD	94%	90%	95%	92%

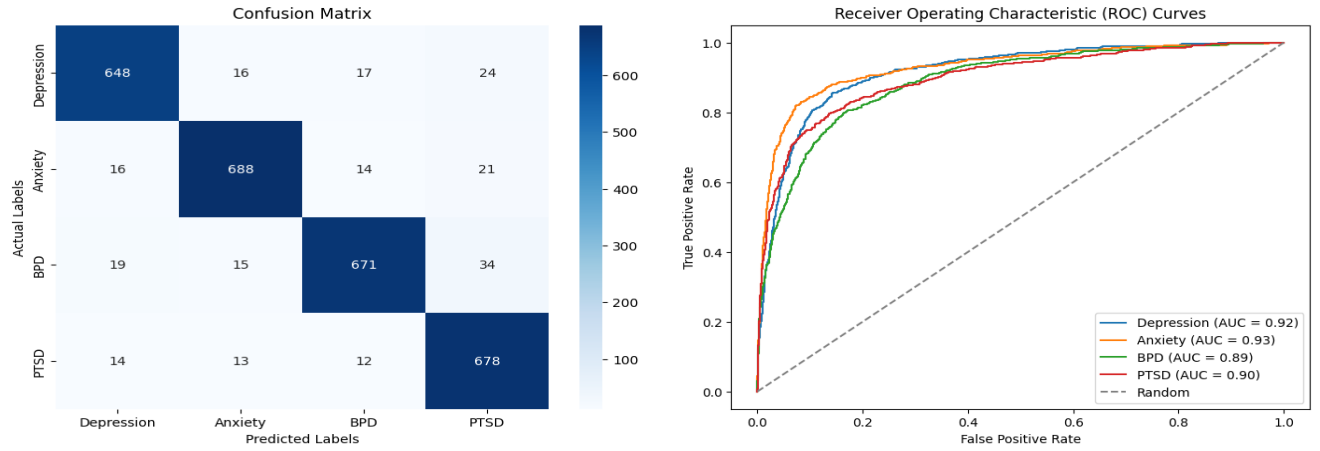
Finally, the F1-score, representing overall performance by combining both precision and recall, was highest for anxiety, with a value of 94%. When evaluated across multiple mental illness categories, anxiety depicted the best performance by maintaining consistently superior performance for all evaluation metrics. Anxiety attained 93% accuracy, 94% precision, 93% recall and 94% F1-score outperforming all other mental illness categories. The other mental illness categories demonstrated fluctuating performance when analyzed in terms of all performance metrics.

An additional dimension for analyzing model performance involves the use of a confusion matrix and an AUC-ROC curve. The confusion matrix provides raw classification



(a) Accuracy. (b) Loss.

FIGURE 5. Training and validation accuracy & loss of the proposed model.



(a) Confusion matrix. (b) AUC-ROC curve.

FIGURE 6. Confusion matrix and AUC-ROC curve for the proposed model.

results on the test data, while the ROC curve calculates AUC values for individual mental illness categories. Fig. 6(a) presents the confusion matrix while Fig. 6(b) displays the ROC curve. The confusion matrix demonstrates that the majority of the results lie on the diagonal, indicating accurate classification by the model. However, there is a minority of samples that deviate from the diagonal. The results show that while the model performs well, there are occasional misclassifications. There are 648 real positives for depression, although 16, 17, and 24 cases are incorrectly labeled as anxiety, BPD, and PTSD respectively. Comparably, anxiety has 688 real positives but has been incorrectly classified as BPD (14), PTSD (21), and depression (16). 671 actual positives for BPD are shown, however, some are incorrectly identified as depression (19), anxiety (15), and PTSD (24). Finally, 678 people actually have PTSD, whereas a small number are incorrectly labeled as BPD (12), depression (14), and anxiety (13). The misclassification in the confusion matrix is attributed to the lack of generalizability in differentiating between mental illness types. BPD seems to be the most challenging to classify correctly, with a comparatively higher rate of confusion with other illnesses. Particularly, there is a significant degree of confusion between BPD and PTSD, which may be caused by shared symptoms or traits between the two disorders. The model generally exhibits good accuracy with just small misclassifications under any conditions. These findings align with the overall model performance discussed in Table 5.

The AUC-ROC curve demonstrates an overview of the model's performance through the area under the curve, affirming that the proposed model has achieved superior performance. The highest AUC value is obtained for anxiety

at 0.93, while depression and PTSD have AUC values of 0.92 and 0.90 respectively. BPD exhibited the lowest performance with a value of 0.89. The results imply that the model has a high degree of accuracy in differentiating between individuals with and without mental illnesses. Correlating this performance with the results presented in Table 5, anxiety emerges as the best-predicted class for mental illness by our model while BPD is identified as the least accurately predicted class. With anxiety having the highest AUC and BPD having the lowest, though still within an acceptable range, these AUC values show that the model functions well under all scenarios. The model appears to be promising overall for screening or diagnosing depression, anxiety, bipolar disorder, and PTSD, according to the AUC-ROC curve data. This analysis indicates that the proposed model is a good fit for real-world uses in determining mental health conditions.

B. ABLATION STUDY

The performance of the proposed model is assessed, including modified versions of hybrid architecture and fine-tuned variants. Specifically, the cascaded CNN model used in both branches of hybrid transformer architecture is subject to analysis with other deep learning models from the NLP domain. This comparison aims to determine which model, when combined with the hybrid architecture, yields the best results for overall mental illness prediction. The objective is to emphasize the unique characteristics and performance traits of every model configuration, highlighting their contributions to the overall performance of modified versions of the model.

1) TRANSFER LEARNING USING BERT, RoBERTa AND DistilBERT

The transfer learning approach analyzes the performance of fine-tuning a pretrained model compared to the proposed model. Fine-tuning a single pretrained model is computationally less intensive than training a hybrid architecture with a cascading of CNN model. The hypothesis of this ablation is that if state-of-the-art results can be achieved by fine-tuning a single model, then designing a complex hybrid architecture may be unnecessary.

The transfer learning results for BERT and its variants are presented in Table 6. The results demonstrate that the RoBERTa model outperforms the BERT and DistilBERT models across all evaluation metrics for all mental illness categories. However, it still falls short of the proposed model by a significant margin. There are only a few instances of over 90% results observed in the RoBERTa fine-tuning. The closest margin of difference is 2%, with the RoBERTa achieving an average accuracy of 90% while, compared to the proposed model's average accuracy of 92%. The differences in precision, recall and F1-score are greater than 3%. These results disprove our hypothesis and validate the recommendation of the proposed model for mental illness prediction.

TABLE 6. Multiclass ablated model results using transfer learning approach for accuracy, precision, recall and F1-Score metrics.

Model	Mental Illness Type	Acc.	Pre.	Rec.	F1-sco.
BERT	Depression	90%	88%	89%	88%
	Anxiety	84%	86%	88%	87%
	BPD	88%	86%	85%	85%
	PTSD	92%	90%	89%	90%
RoBERTa	Depression	92%	91%	90%	90%
	Anxiety	85%	87%	87%	87%
	BPD	89%	89%	90%	90%
	PTSD	94%	92%	93%	92%
DistilBERT	Depression	87%	85%	85%	86%
	Anxiety	81%	83%	83%	82%
	BPD	85%	86%	86%	85%
	PTSD	90%	90%	89%	88%

2) MentalBERT & MelBERT MODELS WITH BiLSTM CASCADING

This ablation study creates a cascaded architecture by channelling the output from hybrid MentalBERT and MelBERT pretrained models to multiple BiLSTM networks for feature extraction. BiLSTM replaces the CNN in the original architecture and acquires contextual feature understanding from the text embeddings generated by both the MentalBERT and MelBERT models. The assumption being that as BiLSTM learns bidirectional contextual information, combining it with attention-based hybrid transformer architecture may produce better results than a CNN model.

The results presented in Table 7 demonstrate the model's performance, indicating comparatively inferior performance to the proposed model by a margin of 4% in terms of average accuracy (see Table 5). The margin for precision, recall and F1-score and even wider than that for accuracy. Although the attention mechanism is intended to complement sequential learning, the results highlight integration issues. The positional encoding in MentalBERT/ MelBERT performs contextual understanding before applying the attention mechanism, and using BiLSTM in a cascading manner creates a duplication effect. This establishes that using BiLSTM as a feature extractor is not a viable approach when cascaded with a hybrid transformer architecture.

TABLE 7. Multiclass ablated model results with BiLSTM cascading for accuracy, precision, recall and F1-score metrics.

Mental Illness Type	Accuracy	Precision	Recall	F1-score
Depression	91%	89%	86%	88%
Anxiety	83%	86%	90%	88%
BPD	87%	85%	86%	85%
PTSD	92%	90%	89%	89%

3) MentalBERT WITH BiLSTM & MelBERT WITH CNN CASCADING

The output from MentalBERT is fed to BiLSTM, and MelBERT is sent to CNN as an input, examining the influence of diverse feature extraction methods within the branches of hybrid transformer architecture. The primary motivation for employing different deep learning models is to enhance the

model's capability to understand text at different levels of granularity.

The class prediction results presented in Table 8 demonstrate model performance across all four mental illness categories. With the introduction of a CNN model in one branch, the results have shown improvement toward the proposed model. The average accuracy difference, which was at least 2% for RoBERTa fine-tuning is now reduced to 1%. However, precision, recall and F1-score still lag behind the proposed model by a significant margin (see Table 5). This supports our hypothesis that CNN model integrates better with the hybrid transformer architecture for feature extraction compared to fine-tuning or a BiLSTM feature extractor.

TABLE 8. Multiclass ablated model results with BiLSTM and CNN cascading for accuracy, precision, recall and F1-score metrics.

Mental Illness Type	Accuracy	Precision	Recall	F1-score
Depression	92%	88%	88%	87%
Anxiety	90%	89%	87%	88%
BPD	88%	86%	87%	86%
PTSD	92%	90%	89%	90%

4) BERT AND DistilBERT MODELS WITH CNN CASCADING

In this ablation study, BERT and DistilBERT, generic embedding models, are replaced within the two branches of the hybrid architecture to generate contextual embeddings. The outputs from each branch are fed into the three-layer CNN models for feature extraction.

The results, presented in Table 9, demonstrate a significant decline in performance when domain-specific MentalBERT and MelBERT models are replaced with BERT and DistilBERT models. The proposed model achieves over 90% accuracy across all evaluation metrics for all mental illness categories (see Table 5) while accuracies are dropped to 90% or below when generic embedding models are used. This supports our hypothesis to use domain-specific pretrained models for generating embedding, as the ablated model demonstrates inferior performance with generic embeddings.

TABLE 9. Multiclass ablated model results with BERT and DistilBERT hybrid architecture for accuracy, precision, recall and F1-score metrics.

Mental Illness Type	Accuracy	Precision	Recall	F1-score
Depression	88%	86%	88%	86%
Anxiety	90%	88%	87%	86%
BPD	85%	85%	86%	86%
PTSD	90%	88%	88%	88%

5) BERT & RoBERTa MODELS WITH CNN CASCADING

This study examines the impact of substituting MentalBERT and MelBERT models with BERT and RoBERTa pretrained models within a hybrid transformer architecture. The cascaded part of the model remains unchanged from the proposed model.

The ablated model displayed subpar results (Table 10) compared to the proposed model (Table 5), with the majority of evaluation metrics falling below 90%, except for a few instances in accuracy metric. Precision, recall and F1-score remain below 90% for all mental illness categories. Despite performing worse than the proposed model, this variant still achieves marginally better results than the architecture using BERT and DistilBERT models. However, the overall conclusion is that replacing BERT and RoBERTa with MentalBERT and MelBERT did not yield superior results.

TABLE 10. Multiclass ablated model results with BERT and RoBERTa hybrid architecture for accuracy, precision, recall and F1-score metrics.

Mental Illness Type	Accuracy	Precision	Recall	F1-score
Depression	90%	88%	87%	86%
Anxiety	91%	88%	89%	88%
BPD	86%	86%	86%	85%
PTSD	90%	88%	87%	87%

6) DUAL-BRANCH BERT MODELS WITH CNN CASCADING

This variant analyses the impact of incorporating BERT_{Base} generic embedding models in each branch of the hybrid transformer architecture. The outputs of the transformer architecture are still fed to the three-layer CNN model for feature extraction.

The results of this ablated version, presented in Table 11, show inferior performance compared to the proposed model (Table 5) and are relatively similar to the BERT and DistilBERT variant. With one exception, all metrics remain below 90% for all mental illness categories. This model again highlights the significance of domain-specific embedding models for mental illness prediction rather than relying on generic embedding models.

TABLE 11. Multiclass ablated model results with Dual-BERT architecture for accuracy, precision, recall and F1-score metrics.

Mental Illness Type	Accuracy	Precision	Recall	F1-score
Depression	88%	87%	88%	87%
Anxiety	90%	86%	87%	87%
BPD	85%	87%	84%	86%
PTSD	89%	89%	85%	86%

7) IMPACT OF CNN VARIANTS ON MODEL PERFORMANCE

This extensive analysis is conducted to evaluate the impact of the varied CNN architecture on the overall performance of the proposed model. Different configurations are tested, including varying the number of CNN layers (1)-(5), and filter sizes (3×3, 5×5), while keeping the number of neurons in the fully connected layers the same as in the proposed model, i.e., 128.

Table 12 compares the performance of CNN models with different numbers of layers. The 4-layer CNN model gets the highest accuracy score of 91%, slightly outperforming the 5-layer CNN model, which has an accuracy score of 90%. This indicates that adding more layers does not always result in significant improvements. The reason is that deeper structures have more filters in each layer, which can capture

TABLE 12. Overall ablated model performance of the proposed model with varying CNN layers.

Configuration	Accuracy	Precision	Recall	F1-score
CNN: 1 layer, FC: 128	81%	83%	83%	82%
CNN: 2 layers, FC: 128	88%	89%	88%	87%
CNN: 4 layers, FC: 128	91%	90%	91%	91%
CNN: 5 layers, FC: 128	90%	91%	90%	90%

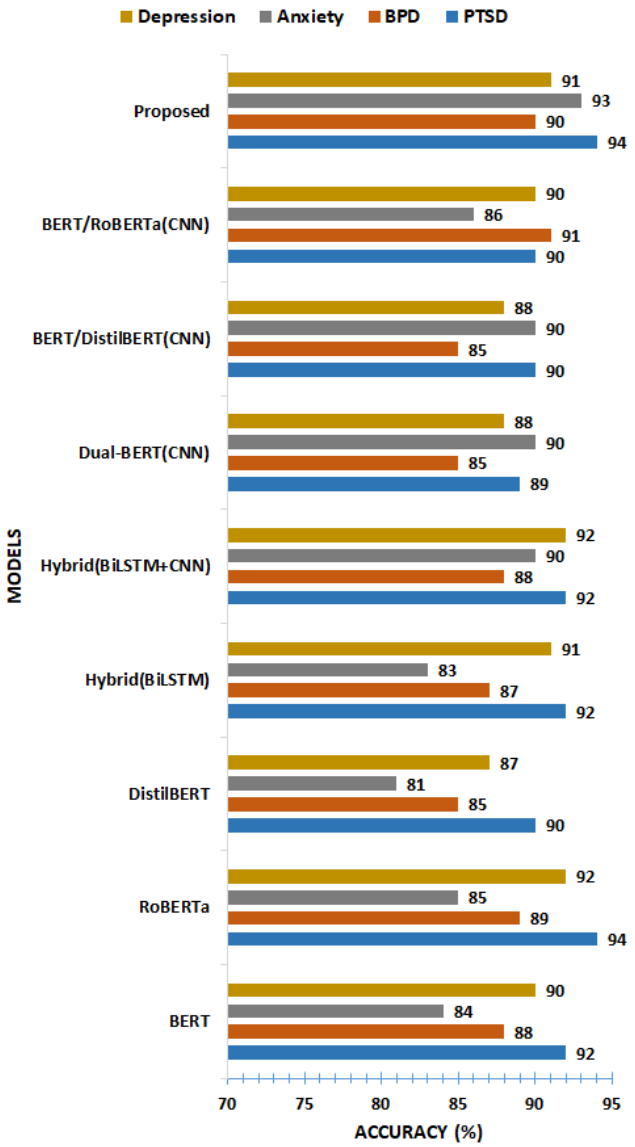


FIGURE 7. Multiclass accuracy comparison of the ablation study results with the proposed model.

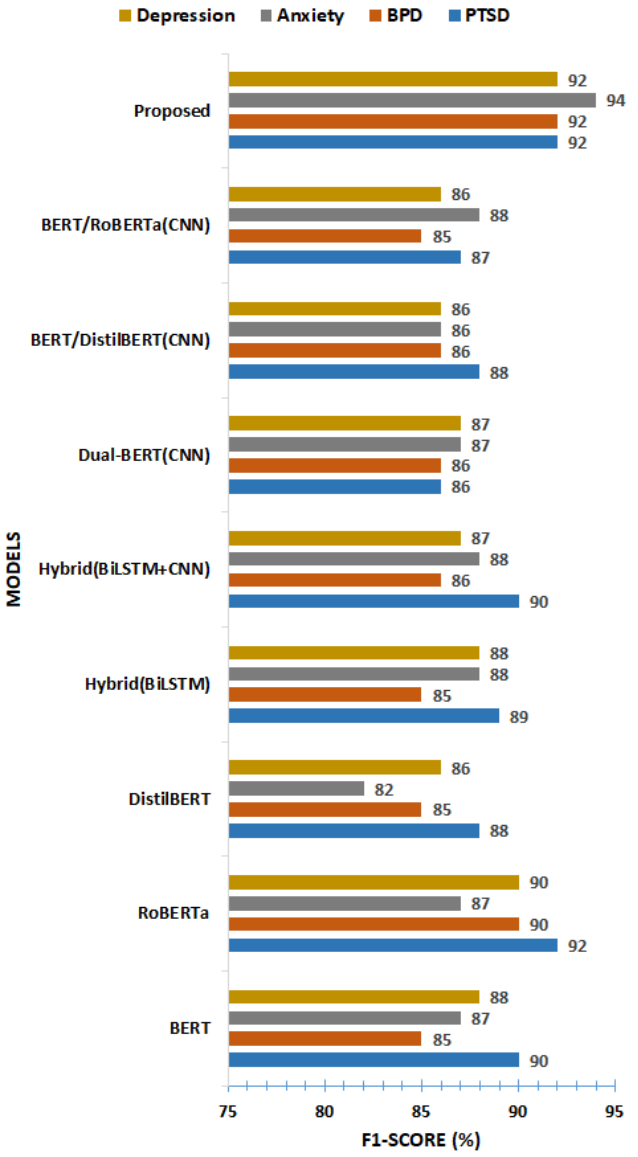


FIGURE 8. Mutliclass F1-score comparison of the ablation study results with the proposed model.

more varied but irrelevant features. This can complicate the learning process and potentially decrease the model's accuracy. Conversely, the 2-layer CNN model fails to extract sufficient relevant features, resulting in a lower performance of 88%. Therefore, the 3-layer CNN model (the proposed model) strikes a balance between complexity and extracting useful features, highlighting the importance of optimizing architecture for robust performance in classification tasks (see Table 5).

C. COMPARATIVE ANALYSIS

The multiclass accuracy comparison between the proposed model and the ablation study results is provided in Fig. 7. The analysis reveals that the proposed model consistently achieved balanced accuracies ranging from 90% to 94% across all mental illness types. The benchmark models, including hybrid (BiLSTM+CNN), BERT/RoBERTa(CNN) and RoBERTa, displayed competitive accuracies ranging from 85% to 94%, surpassing the proposed model's performance for depression and BPD with a margin of 1%. The remaining models demonstrate subpar performance both in terms of accuracy ranges and performance across individual

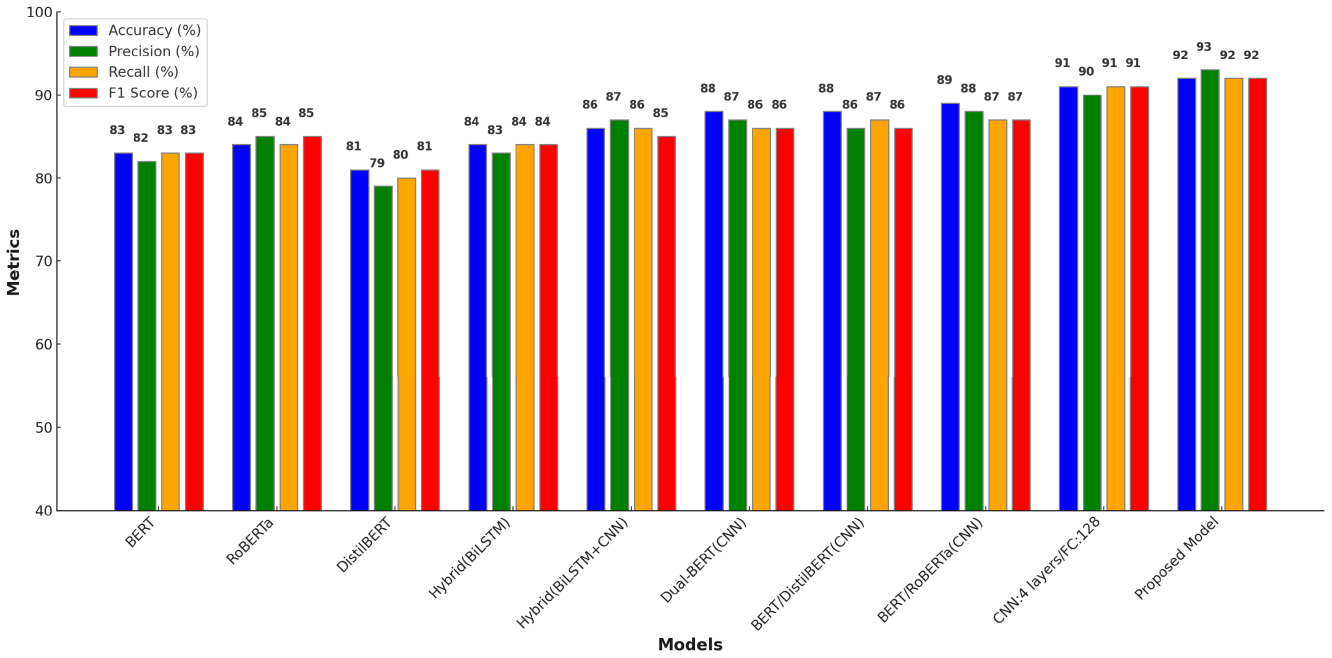


FIGURE 9. Comparison of the ablation study with the proposed model results.

mental illness types. The worst performance is observed with the benchmark models involving generic BERT and DistilBERT, whether fine-tuned or used in a hybrid setup.

Fig. 8 shows the multiclass F1-score comparison between the proposed model and the ablation study results. The proposed model outperforms ablated models by a significant margin, maintaining a consistently high F1-score between 92% to 94% across all mental illness types. The fine-tuned RoBERTa, the best among all the benchmark models fails to surpass the proposed model's performance, with an F1-score ranging from 87% to 92%. The remaining hybrid models and fine-tuned variants exhibit lower F1-scores in the range of 82% to 90% and their performance is suboptimal across all mental illness types compared to the proposed model.

The overall performance for all evaluation metrics is presented in Fig. 9. The results display impressive performance of the proposed architecture with an accuracy of 92%, precision of 93%, recall of 92%, and an F1-score of 92%. In comparison, the best among ablated models achieved an accuracy beyond 91%, prediction of 90%, recall of 91% and F1-score of 91%. These results were obtained using the proposed MentalBERT and MelBERT architecture with four CNN layers. However, an increase in the number of layers beyond four started to degrade the results as shown in Table 12. Among fine-tuned models, the closest in performance is the RoBERTa, with accuracy, precision, recall, and F1-score of 84%, 85%, 84%, and 85%, respectively. The remarkable performance of the proposed model marks a paradigm shift in the field of mental disorder classification. By strategically combining the strengths of various architectures, the proposed model surpasses existing models, attaining a deeper understanding of mental health text.

In the context of overall mental illness classification, our comparative research yields valuable insights into the performance of different models. Fig. 9 shows that while fine-tuned BERT, RoBERTa, and DistilBERT exhibit reasonable accuracy and precision, they fall short of fully capturing the subtleties present in mental health data. A significant improvement is observed with the introduction of the hybrid BERT/RoBERTa architecture cascaded with CNN in both branches, leveraging the collaborative potential of BERT and RoBERTa models. The hybrid architecture incorporating both BiLSTM and CNN in separate branches, demonstrates the increased effectiveness of convolutional neural networks in feature extraction. However, in this performance comparison, the proposed model emerges as the superior performer, surpassing all other models.

VI. DISCUSSION

The current models for diagnosing and classifying mental illnesses demonstrate suboptimal performance as they often rely on traditional embedding techniques such as BOW, TF-IDF, and Glove, or on generic language models like BERT, RoBERTa, and DistilBERT for generating text embeddings. These models struggle with understanding contextual information and comprehending metaphorical expressions in posts of person with a psychiatric disability patients. These challenges primarily arise due to their pretraining on general English language corpus, lacking a specific focus on mental health-related language and nuances.

Our study investigates the use of domain-specific pretrained language models, specifically MentalBERT and MelBERT, for predicting mental illness through the analysis of social media text. The proposed hybrid transformer architecture is cascaded with CNN models for feature extraction. The CNN architecture effectively extracts deep features from the text data, complementing the semantic understanding provided by the MelBERT and MentalBERT language models. The model is simulated on a representative dataset acquired from multiple sources. Additionally, we perform ablation experiments by adjusting the configurations of the proposed architecture for comparative analysis.

The analysis provides valuable insights into the proposed model's effectiveness in classifying mental illness types. It validates the suitability of hybrid transformer architectures for text classification, going beyond the conventional approach of fine-tuning pretrained models. The validation of CNN as a proficient feature extractor in the NLP domain further supports our findings. Our novel architecture marks a significant advancement in the categorization of mental illnesses, outperforming existing models by skillfully combining the benefits of various techniques. This indicates a more profound understanding of mental health literature, enhancing its ability to recognize and classify mental health issues.

VII. LIMITATIONS

The proposed hybrid transformer design shows great promise for identifying mental disorders through text analysis. However, the study is constrained by certain limitations. The model's effectiveness has only been assessed on a small set of data, which may not capture the full spectrum of mental health disorders and language patterns present in wider populations. Moreover, the model's accuracy may differ when applied to other datasets, underscoring the necessity for additional testing on larger and more diverse data sets for validation. Future studies should address these issues to confirm the reliability and applicability of the model.

VIII. CONCLUSION AND FUTURE WORK

This paper presents a comprehensive analysis of the classification of mental illnesses through the use of advanced deep learning techniques. Using pretrained state-of-the-art BERT architectures such as MentalBERT and MelBERT, we analyzed the nuances of social media language to find patterns similar to various mental health problems. By gaining an in-depth understanding of contextual semantics, these transformer-based architectures enabled our model to detect minute linguistic cues associated with mental health. A significant novelty in this work is the hybridization of MentalBERT and MelBERT models, in an attempt to understand the intricate relationships among diverse linguistic expressions present in the textual data. The model performance was further enhanced by cascading pretrained models with convolutional neural networks (CNNs). This hybrid architecture created new opportunities for text representation by combining sequential and hierarchical feature extraction processes. The results demonstrate superior performance of the proposed model with higher values of accuracy, precision, recall and F1-score, which highlight proposed model capacity to identify minor patterns suggestive of mental health conditions. Beyond its contribution to mental health analysis, this research emphasizes how important it is to use hybrid deep learning architectures for sophisticated text understanding in the natural language processing (NLP) domain.

Future research on hybrid models that combine textual models for mental health classification with visual data like facial expressions holds great potential. By taking into account both linguistic and visual indicators, the multimodal approach can further enhance predictions about the mental state of users. The dataset preparation for multimodal mental illness detection is another future research work dimension.

REFERENCES

[1] P. W. Corrigan and K. A. Kosyluk, "Mental illness stigma: Types, constructs, and vehicles for change," in *The Stigma of Disease and Disability: Understanding Causes and Overcoming Injustices*, P. W. Corrigan, Ed., Washington, DC, USA: American Psychological Association, 2014, pp. 35–56.

[2] Y. Gan, H. Huang, X. Wu, and M. Meng, "What doesn't kill us makes us stronger: Insights from neuroscience studies and molecular genetics," *Current Opinion Behav. Sci.*, vol. 59, Oct. 2024, Art. no. 101431.

[3] C. Zhu, "Computational intelligence-based classification system for the diagnosis of memory impairment in psychoactive substance users," *J. Cloud Comput.*, vol. 13, no. 1, p. 119, Jun. 2024.

[4] T. A. Ghebreyesus. (2022). *World Mental Health Report: Transforming Mental Health for All*. Accessed: Feb. 4, 2024. [Online]. Available: https://iris.who.int/bitstream/handle/10665/356119/9789240049338-eng.pdf?sequence=1

[5] P. Bower, S. Knowles, P. A. Coventry, N. Rowland, and C. C. M. D. Group, "Counselling for mental health and psychosocial problems in primary care," *Cochrane Database Systematic Rev.*, vol. 2011, no. 9, Sep. 1996, Art. no. CD001025.

[6] W. N. Stone, "Treatment of the chronically mentally ill: An opportunity for the group therapist," *Int. J. Group Psychotherapy*, vol. 41, no. 1, pp. 11–22, Jan. 1991.

[7] C. Zhu, "Research on emotion recognition-based smart assistant system: Emotional intelligence and personalized services," *J. Syst. Manage. Sci.*, vol. 13, no. 5, pp. 227–242, 2023.

[8] L. Wu, Y. Long, C. Gao, Z. Wang, and Y. Zhang, "MFIR: Multimodal fusion and inconsistency reasoning for explainable fake news detection," *Inf. Fusion*, vol. 100, Dec. 2023, Art. no. 101944.

[9] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proc. Int. AAAI Conf. Web Social Media*, 2020, vol. 7, no. 1, pp. 128–137.

[10] L. Wang, H. Liu, and T. Zhou, "A sequential emotion approach for diagnosing mental disorder on social media," *Appl. Sci.*, vol. 10, no. 5, p. 1647, Mar. 2020.

[11] V. K. Prasad, A. Verma, P. Bhattacharya, S. Shah, S. Chowdhury, M. Bhavsar, S. Aslam, and N. Ashraf, "Revolutionizing "Revolutionizing healthcare: A comparative insight into deep learning's role in medical imaging," *Sci. Rep.*, vol. 14, no. 1, pp. 1–39, Dec. 2024.

[12] H. Khan, N. Javaid, T. Bashir, M. Akbar, N. Alrajeh, and S. Aslam, "Heart disease prediction using novel ensemble and blending based cardiovascular disease detection networks: EnsCVDD-net and BlCVDD-net," *IEEE Access*, vol. 12, pp. 109230–109254, 2024.

[13] H. Pan, Y. Wang, Z. Li, X. Chu, B. Teng, and H. Gao, "A complete scheme for multi-character classification using EEG signals from speech imagery," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 8, pp. 2454–2462, Aug. 2024.

[14] Y. Su, X. Tian, R. Gao, W. Guo, C. Chen, C. Chen, D. Jia, H. Li, and X. Lv, "Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis," *Comput. Biol. Med.*, vol. 145, Jun. 2022, Art. no. 105409.

[15] Z. Zhou, X. Zhou, H. Qi, N. Li, and C. Mi, "Near miss prediction in commercial aviation through a combined model of grey neural network," *Expert Syst. Appl.*, vol. 255, Dec. 2024, Art. no. 124690.
[16] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Health Inf. Sci. Syst.*, vol. 6, no. 1, pp. 1–12, Dec. 2018.
[17] L. Yin, L. Wang, S. Lu, R. Wang, Y. Yang, B. Yang, S. Liu, A. AlSanad, S. A. AlQahtani, Z. Yin, X. Li, X. Chen, and W. Zheng, "Convolution-transformer for image feature extraction," *Comput. Model. Eng. Sci.*, vol. 141, no. 1, pp. 87–106, 2024.
[18] H. Kour and M. K. Gupta, "An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM," *Multimedia Tools Appl.*, vol. 81, no. 17, pp. 23649–23685, Jul. 2022.
[19] I. Ameer, M. Arif, G. Sidorov, H. Gòmez-Adorno, and A. Gelbukh, "Mental illness classification on social media texts using deep learning and transfer learning," 2022, *arXiv:2207.01012*.
[20] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Detection and classification of mental illnesses on social media using RoBERTa," 2020, *arXiv:2011.11226*.
[21] U. Yadav and A. K. Sharma, "A novel automated depression detection technique using text transcript," *Int. J. Imag. Syst. Technol.*, vol. 33, no. 1, pp. 108–122, Jan. 2023.
[22] T. Tran and R. Kavuluru, "Predicting mental conditions based on," *J. Biomed. Informat.*, vol. 75, pp. S138–S148, Aug. 2017.
[23] W. Zheng, G. Gong, J. Tian, S. Lu, R. Wang, Z. Yin, X. Li, and L. Yin, "Design of a modified transformer architecture based on relative position coding," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, p. 168, Oct. 2023.
[24] N. Chen and J. Pan, "The causal effect of delivery volume on severe maternal morbidity: An instrumental variable analysis in Sichuan, China," *BMJ Global Health*, vol. 7, no. 5, May 2022, Art. no. e008428.
[25] Q. Wang, Q. Jiang, Y. Yang, and J. Pan, "The burden of travel for care and its influencing factors in China: An inpatient-based study of travel time," *J. Transp. Health*, vol. 25, Jun. 2022, Art. no. 101353.
[26] M. A. Rahaman, J. Chen, Z. Fu, N. Lewis, A. Iraji, and V. D. Calhoun, "Multi-modal deep learning of functional and structural neuroimaging and genomic data to predict mental illness," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 3267–3272.
[27] Y. Du, L. Chen, M.-C. Yan, Y.-L. Wang, X.-L. Zhong, C.-X. Xv, Y.-B. Li, and Y. Cheng, "Neurometabolite levels in the brains of patients with autism spectrum disorders: A meta-analysis of proton magnetic resonance spectroscopy studies (N = 1501)," *Mol. Psychiatry*, vol. 28, no. 7, pp. 3092–3103, Jul. 2023.
[28] B. Kholifah, I. Syarif, and T. Badriyah, "Mental disorder detection via social media mining using deep learning," *Kinetik, Game Technol., Inf. Syst., Comput. Netw., Comput., Electron., Control*, vol. 5, pp. 309–316, Nov. 2020.
[29] A. Stringaris, "What is depression?" *J. Child Psychol. Psychiatry*, vol. 58, no. 12, pp. 1287–1289, 2017.
[30] M. G. Craske, S. L. Rauch, R. Ursano, J. Prenoveau, D. S. Pine, and R. E. Zinbarg, "What is an anxiety disorder?" *Focus*, vol. 9, no. 3, pp. 369–388, 2011.
[31] M. Bohus, J. Stoffers-Winterling, C. Sharp, A. Krause-Utz, C. Schmahl, and K. Lieb, "Borderline personality disorder," *The Lancet*, vol. 398, no. 10310, pp. 1528–1540, 2021.
[32] M. Khazbak, Z. Wael, Z. Ehab, M. Gerorge, and E. Eliwa, "MindTime: Deep learning approach for borderline personality disorder detection," in *Proc. Int. Mobile, Intell., Ubiquitous Comput. Conf. (MIUCC)*, May 2021, pp. 337–344.
[33] K. Nova, "Machine learning approaches for automated mental disorder classification based on social media textual data," *Contemp. Issues Behav. Social Sci.*, vol. 7, no. 1, pp. 70–83, 2023.
[34] R. Yehuda, C. W. Hoge, A. C. McFarlane, E. Vermetten, R. A. Lanius, C. M. Nievergelt, S. E. Hobfoll, K. C. Koenen, T. C. Neylan, and S. E. Hyman, "Post-traumatic stress disorder," *Nature Rev. Disease Primers*, vol. 1, no. 1, pp. 1–22, Oct. 2015.
[35] G. Coppersmith, C. Harman, and M. Dredze, "Measuring post traumatic stress disorder in Twitter," in *Proc. Int. AAAI Conf. Web Social Media*, May 2014, vol. 8, no. 1, pp. 579–582.
[36] U. Rashida and K. Suresh Kumar, "Social media mining to detect mental health disorders using machine learning," in *Sentiment Analysis and Deep Learning*. Cham, Switzerland: Springer, 2023, pp. 923–930.
[37] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: An integrative review," *Current Opinion Behav. Sci.*, vol. 18, pp. 43–49, Dec. 2017.
[38] H.-H. Shuai, C.-Y. Shen, D.-N. Yang, Y. C. Lan, W.-C. Lee, P. S. Yu, and M.-S. Chen, "A comprehensive study on social network mental disorders detection via online social media mining," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 7, pp. 1212–1225, Jul. 2018.
[39] A. Priya, S. Garg, and N. P. Tigga, "Predicting anxiety, depression and stress in modern life using machine learning algorithms," *Proc. Comput. Sci.*, vol. 167, pp. 1258–1267, Jul. 2020.
[40] I. Fatima, B. U. D. Abbasi, S. Khan, M. Al-Saeed, H. F. Ahmad, and R. Mumtaz, "Prediction of postpartum depression using machine learning techniques from social media text," *Expert Syst.*, vol. 36, no. 4, p. 12409, Aug. 2019.
[41] J. Chung and J. Teo, "Single classifier vs. Ensemble machine learning approaches for mental health prediction," *Brain Informat.*, vol. 10, no. 1, pp. 1–10, Dec. 2023.
[42] V. Tejaswini, K. Sathya Babu, and B. Sahoo, "Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 23, no. 1, pp. 1–20, Jan. 2024.
[43] M. Aragon, A. P. Lopez Monroy, L. Gonzalez, D. E. Losada, and M. Montes, "DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, ON, Canada: Association for Computational Linguistics, 2023, pp. 15305–15318. [Online]. Available: https://aclanthology.org/2023.acl-long.853
[44] P. Seth and M. Agarwal, "UATTA-EB: Uncertainty-aware test-time augmented ensemble of BERTs for classifying common mental illnesses on social media posts," 2023, *arXiv:2304.04539*.
[45] M. Kabir, T. Ahmed, M. B. Hasan, M. T. R. Laskar, T. K. Joarder, H. Mahmud, and K. Hasan, "DEPTWEET: A typology for social media texts to detect depression severities," *Comput. Hum. Behav.*, vol. 139, Feb. 2023, Art. no. 107503.
[46] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang, "Mental-LLM: Leveraging large language models for mental health prediction via online text data," 2023, *arXiv:2307.14385*.
[47] W. R. D. Santos, R. L. de Oliveira, and I. Paraboni, "SetembroBR: A social media corpus for depression and anxiety disorder prediction," *Lang. Resour. Eval.*, vol. 58, no. 1, pp. 273–300, Mar. 2024.
[48] I. Tavchioski, M. Robnik-Šikonja, and S. Pollak, "Detection of depression on social networks using transformers and ensembles," 2023, *arXiv:2305.05325*.
[49] E. Martinez, J. Cuadrado, J. C. Martinez-Santos, D. Peña, and E. Puertas, "Automated depression detection in text data: Leveraging lexical features, phonesthemes embedding, and roberta transformer model," in *Proc. CEUR Workshop*, 2023, pp. 1–14.
[50] 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. (LREC)*. Marseille, France: European Language Resources Association, 2022, pp. 7184–7190. [Online]. Available: https://aclanthology.org/2022.lrec-1.778
[51] M. Niu, K. Chen, Q. Chen, and L. Yang, "HCAG: A hierarchical context-aware graph attention model for depression detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4235–4239.
[52] P. Arora and P. Arora, "Mining Twitter data for depression detection," in *Proc. Int. Conf. Signal Process. Commun. (ICSC)*, Mar. 2019, pp. 186–189.
[53] A.-S. Uban, B. Chulvi, and P. Rosso, "An emotion and cognitive based analysis of mental health disorders from social media data," *Future Gener. Comput. Syst.*, vol. 124, pp. 480–494, Nov. 2021.
[54] 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.
[55] N. K. Iyortsuun, S.-H. Kim, M. Jhon, H.-J. Yang, and S. Pant, "A review of machine learning and deep learning approaches on mental health diagnosis," *Healthcare*, vol. 11, no. 3, p. 285, Jan. 2023.
[56] M. Bhuvaneswari and V. L. Prabha, "A deep learning approach for the depression detection of social media data with hybrid feature selection and attention mechanism," *Expert Syst.*, vol. 40, no. 9, p. 13371, Nov. 2023.

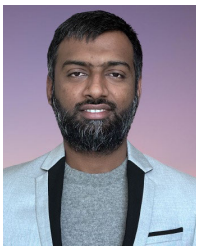
[57] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Sci. Rep.*, vol. 10, no. 1, p. 11846, Jul. 2020.
[58] G. Rao, Y. Zhang, L. Zhang, Q. Cong, and Z. Feng, "MGL-CNN: A hierarchical posts representations model for identifying depressed individuals in online forums," *IEEE Access*, vol. 8, pp. 32395–32403, 2020.
[59] M. Garg, "Multi-class categorization of reasons behind mental disturbance in long texts," *Knowl.-Based Syst.*, vol. 276, Sep. 2023, Art. no. 110742, doi: 10.1016/j.knosys.2023.110742.
[60] L. Bendebane, Z. Laboudi, A. Saighi, H. Al-Tarawneh, A. Ouannas, and G. Grassi, "A multi-class deep learning approach for early detection of depressive and anxiety disorders using Twitter data," *Algorithms*, vol. 16, no. 12, p. 543, Nov. 2023.
[61] M. Choi, S. Lee, E. Choi, H. Park, J. Lee, D. Lee, and J. Lee, "MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories," 2021, *arXiv:2104.13615*.

ADNAN KARAMAT received the bachelor's degree in computer science from Pir Mehr Ali Shah Arid Agricultural University, Rawalpindi, in 2019, and the master's degree in computer science from COMSATS University Islamabad, in 2024. With a passion for research and a strong academic foundation, he has honed a diverse skill set showcasing expertise in research methodology, advanced technical writing, critical analytical thinking, robust data analysis, and adept problem-solving capabilities. His research interests include artificial intelligence, deep learning, machine learning, natural language processing, and computer vision.

MUHAMMAD IMRAN received the degree in software engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2006, and the master's degree in software engineering and the Ph.D. degree in computer science from the University of Southampton, U.K., in 2009 and 2015, respectively. He was a Lecturer with COMSATS University Islamabad (CUI), Islamabad, Pakistan, from 2007 to 2008. He is currently an Assistant Professor with the Department of Computer Science, CUI. His research interests include artificial intelligence, machine learning, natural language processing, and the semantic web.

MUHAMMAD USMAN YASEEN is currently an Assistant Professor with COMSATS University Islamabad (CUI), Islamabad, Pakistan. His current research interests include video analytics, big data analysis, machine learning, and distributed systems.

RASOOL BUKHSH is currently an Assistant Professor at COMSATS University Islamabad (CUI), Islamabad, Pakistan. His current research interests include smart grids, energy management, video analytics, big data analysis, machine learning, and distributed systems.



SHERAZ ASLAM received the B.S. degree in computer science from Bahauddin Zakariya University (BZU), Multan, Pakistan, in 2015, the M.S. degree in computer science with a specialization in energy optimization in the smart grid from COMSATS University Islamabad (CUI), Islamabad, Pakistan, in 2018, and the Ph.D. degree in computer engineering and informatics from Cyprus University of Technology (CUT), Limassol, Cyprus, under the supervision of Dr. Herodotos Herodotou. He was a Research Associate with Dr. Nadeem Javaid during the M.S. degree with CUI. He is currently a Postdoctoral Researcher with the DICL Research Laboratory, CUT, where he is also a part of a European Union-funded research project named as aerOS. Previously, he was involved in several EU-funded research projects, i.e., STEAM and MARISense. He has authored more than 100 research publications in ISI-indexed international journals and conferences, including the IEEE Internet of Things Journal, *Renewable and Sustainable Energy Reviews*, and *Electric Power Systems Research*. His research interests include data analytics, generative adversarial networks, wireless networks, smart grids, and cloud computing. He is constantly looking for collaboration opportunities with professors and students from different universities around the globe.

NOUMAN ASHRAF (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from COMSATS University Islamabad, Pakistan, and the Ph.D. degree in electrical engineering from Frederick University, Cyprus, under the Erasmus Mundus Scholarship Program. He was with the Turku University of Applied Sciences, Finland, the TSSG, Waterford Institute of Technology, Ireland, and the University of Cyprus. He is currently with Technological University Dublin, Ireland. His research interests include the application of control theory for the management of emerging networks with applications in the Internet of Things, 5G and beyond communication networks, electric vehicles, metamaterial-based beam steering, wireless sensor networks, smart grids, network resilience, and power load modeling.