IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ADVANCED MULTICLASS MENTAL ILLNESS PREDICTION USING A HYBRID MENTALBERT–MELBERT TRANSFORMER ARCHITECTURE

[1] Mr. K. Raveendra Chaitanya, [2] Dr. Lakshmana Rao Battarusetty

[1] Assistant Professor, [2] Professor

[*]*Department of Computer Science and Engineering*

Geethanjali Institute of Science and Technology (Autonomous), Nellore, Andhra Pradesh, India

Abstract: Mental illness prediction from text uses Natural Language Processing (NLP) and deep learning to identify mental health conditions from written content such as social media posts. Traditional classification models, including Logistic Regression and Naive Bayes, often struggle to understand the emotional context and metaphorical expressions commonly found in mental health-related communication, since they rely on simple feature extraction methods such as TF-IDF and assume independence between words. To address this limitation, this paper proposes a hybrid transformer-based architecture that combines MentalBERT, MelBERT, and Convolutional Neural Networks (CNN) for multiclass mental illness prediction. MentalBERT captures domain-specific language patterns associated with mental health, MelBERT interprets metaphorical and figurative expressions, and CNN layers extract deeper spatial features from the resulting embeddings before classification. The model is trained and evaluated on a balanced dataset of 40,000 social media text samples drawn from Reddit communities and a PTSD corpus, covering four classes: Depression, Anxiety, Borderline Personality Disorder (BPD), and Post-Traumatic Stress Disorder (PTSD). Experimental results show that the proposed hybrid model achieves a test accuracy of 92%, precision of 93%, recall of 92%, and an F1-score of 92%, with a best validation accuracy of 93.63% recorded during training, outperforming baseline models such as Logistic Regression, Naive Bayes, and standalone transformer architectures. The system is deployed through a web-based interface that returns the predicted condition together with a confidence score and class probability distribution, supporting its potential use in social media monitoring and clinical text analysis for early detection of mental health risks.

Index Terms - Mental Illness Prediction, MentalBERT, MelBERT, Convolutional Neural Network, Transformer Architecture, Natural Language Processing, Multiclass Classification, Social Media Text Analysis, Deep Learning.

I. INTRODUCTION

Mental health has become a critical concern in modern society, with a marked rise in psychological disorders such as depression, anxiety, borderline personality disorder (BPD), and Post-Traumatic Stress Disorder (PTSD). With the rapid growth of social media platforms, individuals increasingly express their emotions, thoughts, and mental states through textual content, opening new opportunities for leveraging Natural Language Processing (NLP) and Artificial Intelligence (AI) to detect mental health conditions automatically. Traditional approaches to mental illness detection rely on manual clinical diagnosis, which is time-consuming, subjective, and difficult to scale. Recent advancements in machine learning and deep learning have enabled automated systems to analyze large volumes of textual data and identify mental health patterns; however, existing models often struggle to capture deep contextual meaning and the metaphorical expressions that are common in social media text.

Mental health disorders affect millions of people worldwide, yet many cases remain undiagnosed due to lack of awareness, social stigma, and limited access to professional healthcare. Early detection plays a crucial role in preventing severe outcomes and improving quality of life, but traditional diagnostic methods require expert involvement and are often slow and expensive. Although existing machine learning and deep learning models have shown promising results, they have limitations in understanding complex linguistic patterns, especially metaphorical expressions commonly used in mental health discussions. This motivates the need for a more advanced and intelligent system that improves prediction accuracy while remaining scalable and accessible through a web-based platform.

Despite progress in AI-based mental illness detection, several challenges remain. Traditional machine learning models fail to capture deep contextual relationships in text, while generic transformer models such as BERT lack domain-specific understanding of mental health language and struggle to interpret figurative expressions in social media posts. Multiclass classification of mental illnesses is further complicated by overlapping symptoms and linguistic similarities among conditions, and the lack of user-friendly systems for real-time prediction limits practical usability. To address these gaps, this paper proposes an Advanced Multiclass Mental Illness Prediction System using a Hybrid MentalBERT–MelBERT Transformer Architecture that combines domain-specific transformer models with CNN-based feature extraction, and is deployed through a web-based application for real-time prediction and user interaction.

The remainder of this paper is organized as follows: Section II presents a review of related work; Section III describes the proposed methodology; Section IV discusses the implementation details; Section V presents the results and discussion; and Section VI concludes the paper with directions for future work.

II. LITERATURE SURVEY

Predicting mental illness from social media text has gained significant attention due to the massive increase in user-generated content on platforms such as Reddit and Twitter, combined with rapid advances in NLP. Researchers have explored a range of techniques, from traditional machine learning approaches such as Support Vector Machines (SVM) to advanced deep learning models such as Recurrent Neural Networks (RNN) and transformer-based models like BERT, each attempting to detect language patterns associated with different mental health conditions. However, many existing approaches still suffer from limited contextual understanding, restricted datasets, and reduced accuracy, motivating the development of hybrid models that combine the strengths of multiple techniques.

An LSTM-based approach has been used to classify the type of mental illness expressed in social media posts, with text preprocessed using stemming, lemmatization, and stop-word removal before feature selection and classification, and its performance compared against other machine learning models on the Reddit Mental Health Dataset. Other work has focused on mitigating data imbalance and poorly labeled data in mental health detection by combining weak classifiers with gradient boosting alongside LSTM and pretrained transformer models such as BERT and RoBERTa, using data resampling, augmentation, and a feature fusion method that combines local features, global features, and TF-IDF representations to improve accuracy, recall, and F1-score.

Closely related to the present work, a hybrid transformer-based model combining MentalBERT and MelBERT with CNN layers has been proposed for multiclass mental illness prediction, where MentalBERT is trained on mental health-related text and MelBERT is trained on metaphorical data to better interpret figurative language; this approach reported an accuracy and F1-score of 92%, outperforming existing models. A multi-task Opinion-Enhanced Hybrid BERT model has also been proposed to jointly perform sentiment analysis and status classification by combining opinion-based features extracted using tools such as TextBlob with BERT's contextual embeddings, CNN layers for local features, and BiGRU layers for sequential patterns, achieving high accuracy in both sentiment and status classification tasks. In a different direction, randomized neural networks such as the Broad Learning System (BLS), Random Vector Functional Link Network (RVFLN), and their kernelized variants have been combined with a TOPSIS-ModCHI feature selection method to detect mental illness from social media text, with the BLS variant achieving strong precision and F1-score even with a reduced feature set.

Table 1: Summary of Existing Mental Illness Detection Approaches

Approach	Technique	Key Limitation
LSTM-based classification	Sequential deep learning on preprocessed Reddit posts	Limited contextual and metaphor understanding
Transfer learning with feature fusion	BERT/RoBERTa with local, global, and TF-IDF feature fusion	Sensitive to class imbalance and label noise
Hybrid MentalBERT + MelBERT + CNN	Domain-specific and metaphor-aware transformer fusion	Computationally demanding to train
Opinion-enhanced hybrid BERT	Multi-task sentiment and status classification with CNN/BiGRU	Focuses on opinion features over deep emotional context
Randomized neural networks (BLS, RVFLN)	TOPSIS-ModCHI feature selection with shallow learners	Limited capacity to model long-range dependencies

Across these studies, traditional machine learning baselines such as Logistic Regression and Naive Bayes remain useful for benchmarking but cannot understand complex language patterns such as sarcasm or metaphors, rely on feature extraction methods like TF-IDF that lose contextual meaning, and perform poorly on imbalanced or long-sequence text. These limitations highlight the continuing need for hybrid architectures that combine domain-specific pretrained language models with metaphor-aware components and convolutional feature extraction, which forms the basis of the methodology proposed in this paper.

III. PROPOSED METHODOLOGY

The proposed system performs multiclass mental illness prediction using a hybrid architecture that combines two domain-adapted transformer models, MentalBERT and MelBERT, with Convolutional Neural Network (CNN) layers for deep feature extraction. The overall pipeline consists of data collection, preprocessing, tokenization, dual-branch transformer embedding generation, CNN-based feature extraction, feature fusion, classification, and deployment through a web-based interface.

A. Dataset Description

The dataset used in this work is collected from multiple mental-health-related sources, including the Reddit communities r/depression, r/anxiety, and r/BPD, along with a HuggingFace PTSD dataset. It consists of 40,000 samples evenly distributed across four classes, Depression, Anxiety, BPD, and PTSD, with 10,000 samples per class, ensuring a balanced dataset for training and evaluation. The average length of the text samples is approximately 127 characters. The original, unbalanced data distribution contained 15,234 Depression, 12,456 Anxiety, 8,932 BPD, and 7,821 PTSD samples, which was subsequently balanced through random downsampling to 10,000 samples per class to avoid bias toward majority categories. The dataset is split into training, validation, and test sets in an 84.5%/8%/7.5% ratio, corresponding to 33,800, 3,200, and 3,000 samples respectively.

B. Data Preprocessing

Data preprocessing prepares the raw social media text for model training and includes text cleaning, normalization, emoticon removal, dataset balancing, dataset splitting, and tokenization. Text cleaning removes unwanted elements such as URLs, HTML tags, special characters, and duplicate entries, while normalization converts text to lowercase and removes extra spacing while preserving meaningful expressions. The dataset is balanced using random downsampling so that all four classes are equally represented, after which it is divided into training, validation, and test sets. Tokenization is performed using the BERT tokenizer, which converts text into numerical representations using special tokens such as [CLS], [SEP], and [PAD], with a maximum sequence length of 512 tokens, preparing the data for input into the transformer models.

C. Model Architecture

The proposed model integrates two transformer branches with parallel CNN feature extractors. MentalBERT is a domain-specific transformer pretrained on a large corpus of mental-health-related text, enabling it to capture subtle cues such as distress, negativity, and psychological states that general-purpose language models often miss. MelBERT is a transformer trained on metaphor-rich data and is used to interpret figurative expressions common in mental health discourse, such as “I feel empty” or “I’m drowning in thoughts.” Both models generate 768-dimensional contextual embeddings, which are largely kept frozen during training to preserve their pretrained domain knowledge while reducing computational complexity.

The embeddings produced by MentalBERT and MelBERT are each passed through CNN feature extractors employing multiple filters of varying kernel sizes (such as 3 and 5) followed by max-pooling, allowing the model to capture local and multi-scale n-gram patterns within the text in addition to the global context already encoded by the transformers. The resulting feature maps from both branches are combined through feature concatenation to form a unified 768-dimensional representation. This fused vector is passed through a dense layer with ReLU activation and dropout regularization to reduce overfitting, and finally through a softmax classification layer that assigns the input text to one of the four mental health categories.

D. Model Training

The model is trained using a batch size of 64, a learning rate of 0.001, and the Adam optimizer, for 40 epochs on a GPU-enabled system. Training and validation accuracy and loss are monitored throughout training to track learning progress. The model achieves a best validation accuracy of 93.63% with a corresponding validation loss of 0.3443, with accuracy increasing steadily and loss decreasing across epochs, indicating stable convergence without significant overfitting.

Table 2: Key Training Hyperparameters

Parameter	Value
Batch size	64
Learning rate	0.001
Optimizer	Adam
Number of epochs	40
Embedding dimension (MentalBERT / MelBERT)	768
Maximum token length	512
Train / Validation / Test split	84.5% / 8% / 7.5%

IV. IMPLEMENTATION

A. Software and Tools

The system is implemented using Python as the primary programming language. Deep learning components, including the transformer branches and CNN layers, are built using TensorFlow and PyTorch, while NLP libraries such as NLTK and Hugging Face Transformers are used for text preprocessing, tokenization, and embedding generation. Scikit-learn is used to implement baseline models such as Logistic Regression and Naive Bayes and to compute evaluation metrics. NumPy and Pandas support data handling, and Matplotlib and Seaborn are used for visualization. Jupyter Notebook and Google Colab are used for experimentation and model training, and the trained model is deployed through a Streamlit-based web interface that supports manual text entry, CSV upload, and sample-case prediction.

B. System Architecture

The system architecture follows a sequential pipeline. Input text first undergoes preprocessing and tokenization using the BERT tokenizer. The tokenized data is passed in parallel through the MentalBERT and MelBERT branches to generate contextual and metaphor-aware embeddings respectively. Each branch's embeddings are further processed by dedicated CNN feature extractors, and the resulting feature vectors are fused through concatenation. The fused representation is passed through dense layers with dropout before a softmax layer produces the final classification across the four mental illness categories, along with a confidence score for the prediction.

C. Testing

A combination of unit and integration testing is used to validate the system. Unit testing verifies individual components such as data preprocessing, tokenization, model prediction, and the web interface independently, while integration testing confirms that the complete pipeline, from raw input text to final prediction output, functions correctly end to end. This testing strategy ensures the reliability of each module before and after integration, reducing the likelihood of failures in deployment.

V. RESULTS AND DISCUSSION

The proposed hybrid MentalBERT–MelBERT model is evaluated on the held-out test set of 3,000 samples using standard classification metrics: accuracy, precision, recall, and F1-score. The model achieves an overall test accuracy of 92%, precision of 93%, recall of 92%, and an F1-score of 92%, which is consistent with the best validation accuracy of 93.63% recorded during training. These results indicate that the hybrid architecture, which jointly leverages domain-specific contextual understanding from MentalBERT, metaphor interpretation from MelBERT, and local pattern extraction from CNN layers, is effective at distinguishing between the four target mental health conditions.

Per-class evaluation shows consistently strong and balanced performance across all categories, with class-wise accuracy in the range of approximately 90% to 94% and F1-scores between 92% and 94% for Depression, Anxiety, BPD, and PTSD. Analysis of the confusion matrix on the 3,000-sample test set shows that the majority of predictions fall along the diagonal, with most classes correctly predicted in roughly 650 to 690 out of around 700 to 750 samples, and only a small number of misclassifications, typically between 12 and 24 samples per class pair. The most frequent confusions occur between classes with overlapping symptom expression, such as Anxiety and PTSD, which is consistent with the linguistic similarity of distress-related language across these conditions.

Table 3: Performance Comparison with Baseline Approaches

Model	Reported Accuracy
Traditional machine learning (e.g., Logistic Regression, Naive Bayes)	~75%
LSTM-based model	~85%
Standalone BERT model	~88%
Proposed Hybrid MentalBERT + MelBERT + CNN	~92%

As summarized in Table 3, the proposed hybrid model outperforms traditional machine learning baselines as well as standalone LSTM and BERT models reported in this study. This improvement is attributed to the combined effect of domain-specific pretraining in MentalBERT, the metaphor-aware embeddings produced by MelBERT, and the additional local feature extraction performed by the CNN layers, which together allow the model to capture both global contextual meaning and fine-grained linguistic patterns that simpler models miss.

The system also supports real-time prediction through a web-based interface, where users can submit text manually, upload a CSV file, or select sample cases. For an example input describing feelings of hopelessness and low motivation, the system predicted the Depression class with a confidence score of 69.6%, with the class probability distribution clearly favoring Depression over the remaining three categories. This demonstrates the practical usability of the system for generating interpretable, real-time predictions together with confidence indicators, which is valuable for applications such as social media monitoring and supporting early detection of mental health risk.

VI. CONCLUSION

This paper presented an Advanced Multiclass Mental Illness Prediction System based on a Hybrid MentalBERT–MelBERT Transformer Architecture for classifying Depression, Anxiety, BPD, and PTSD from social media text. By combining MentalBERT for domain-specific contextual understanding, MelBERT for interpreting metaphorical and figurative expressions, and CNN layers for additional spatial feature extraction, the proposed system addresses key limitations of traditional machine learning models and generic transformer architectures, namely their inability to capture deep emotional context and figurative language in mental health text. Trained on a balanced dataset of 40,000 samples and evaluated on a held-out test set, the model achieves a test accuracy, precision, and F1-score in the range of 92% to 93%, with a best validation accuracy of 93.63%, outperforming traditional baselines such as Logistic Regression and Naive Bayes as well as standalone deep learning models considered in this study.

The developed system is deployed through a web-based interface that enables real-time prediction with confidence scores and class probability visualization, demonstrating its practical applicability for social media monitoring and clinical text analysis in support of early mental health risk detection. Future work may focus on extending the architecture to additional mental health conditions, improving robustness to short or ambiguous text inputs, incorporating explainable AI techniques to enhance interpretability for clinical use, and exploring lightweight or quantized versions of the model to enable deployment on resource-constrained or mobile platforms.

REFERENCES

[1] A. Karamat et al., “A hybrid transformer architecture for multiclass mental illness prediction using social media text,” IEEE Access, vol. 13, pp. 12148–12167, 2024.

[2] K. Hasan et al., “Advancing mental disorder detection: a comparative evaluation of transformer and LSTM architecture on social media text,” 2025.

[3] Y. Zhu and S. Huang, “Social text mental disorder identification framework integrating generative language model and psychological features extraction,” 2025.

[4] N. Absar et al., “Transformer and deep learning-based depression detection from low-resource language,” 2024.

[5] J. Chung and J. Teo, “Mental health prediction using machine learning: Taxonomy, applications, and challenges,” Appl. Comput. Intell. Soft Comput., vol. 2022, 2022.

[6] M. Garg, “Mental health analysis in social media posts: A survey,” Arch. Comput. Methods Eng., vol. 30, pp. 1819–1842, 2023.

[7] M. M. Aldarwish and H. F. Ahmad, “Predicting depression levels using social media posts,” in Proc. IEEE 13th Int. Symp. Autonomous Decentralized Syst. (ISADS), 2017, pp. 277–280.

[8] R. Thorstad and P. Wolff, “Predicting future mental illness from social media: A big-data approach,” Behav. Res. Methods, vol. 51, pp. 1586–1600, 2019.

[9] S. Hussain, A. Nasir, K. Aslam, S. Tariq, and M. F. Ullah, “Predicting mental illness using social media posts and comments,” Int. J. Adv. Comput. Sci. Appl., vol. 11, pp. 607–613, 2020.

[10] G. Ansari, M. Garg, and C. Saxena, “Data augmentation for mental health classification on social media,” arXiv:2112.10064, 2021.

[11] 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, p. 11846, 2020.

[12] 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,” Multimed. Tools Appl., vol. 81, pp. 23649–23685, 2022.

[13] A. Murarka, B. Radhakrishnan, and S. Ravichandran, “Detection and classification of mental illnesses on social media using RoBERTa,” arXiv:2011.11226, 2020.

[14] B. G. Bokolo and Q. Liu, “Deep learning-based depression detection from social media: Comparative evaluation of ML and transformer techniques,” Electronics, vol. 12, p. 4396, 2023.

[15] L. Zhu, S. Hou, and T. Cheng, “Mental health status detection model based on LSTM neural network,” Procedia Comput. Sci., vol. 243, pp. 842–849, 2024.

[16] 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, p. 543, 2023.

[17] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, “MentaLLaMA: Interpretable mental health analysis on social media with large language models,” in Proc. ACM Web Conf., 2024, pp. 4489–4500.

[18] H. He and Y. Ma, Eds., Imbalanced Learning: Foundations, Algorithms, and Applications. Hoboken, NJ, USA: John Wiley & Sons, 2013.