Explainable Mental Health Chat Support using MKO-DEBiGRU with Knowledge Graph and SHAP

P. Nagarathna[1], *Srujana* Thalla[2], *Sujata* Patil[3*], and *Phanindra* Reddy. K[4]

[1]Department of Computer Science, Nitte Meenakshi Institute of Technology, Nitte (Deemed to be University), Bengaluru, India
[2]Department of Computer Science and Engineering (AI&ML), Geetanjali College of Engineering and Technology, Hyderabad, India
[3]Department of Electronics and Communication Engineering, KLE Technological University, Hubli India
[4]Department of Artificial Intelligence and Machine Learning, Ballari Institute of Technology & Management, Ballari, India

Abstract. Recently, mental health chat support has become increasingly important for providing timely assistance and guidance to users seeking psychological assistance. This task becomes more challenging because of unstructured and noisy text inputs, where slang, abbreviations, and inconsistent phrasing limit the reliability of the responses. However, traditional models struggle to maintain accuracy and offer explainable reasoning in these scenarios. Hence, this study proposes a Medical Knowledge Graph integrated orthogonal independent Deep Bidirectional Gated Recurrent Unit (MKO-DEBiGRU) method for mental health text classification. To enhance interpretability, a Medical Knowledge Graph is used to map the predictions to domain knowledge, and SHapley additive explanations (SHAP)-based Explainable AI (XAI) is integrated to provide feature-level insights into model decisions. The proposed model, MKO-DEBiGRU, is evaluated using the Anxiety, ADHD, Bipolar, and Depression datasets, where preprocessing includes removing stop words, punctuation, hashtags, and special characters to improve text quality. The experimental results show that the proposed MKO-DEBiGRU method achieves 99.98% accuracy, significantly outperforming existing BiLSTM models in terms of classification performance.

1 Introduction

A significant burden on individuals and societies worldwide is imposed by mental health issues According to a recent report, over 20% of adults experience some form of mental disorder at least once in their lifetime [1]. The growing impact of digitalization on healthcare, Artificial Intelligence (AI) emerging as a tool transforming mental health research and practice [2]. With the rapid development of AI and Cyber Physical Systems (CPS), there is an increasing demand for interpretability and transparency of AI technologies in practical applications [3]. The Support, Management, Individual, Learning Enablement (SMILE) platform is a comprehensive AI-enabled framework that addresses clinician burnout, data privacy, and integration of therapeutic modules into daily healthcare workflows [4]. To identify potential biases affecting trust in AI for psychological advice and provide a qualified assessment of AI-generated content by focusing on clinicians' evaluations [5]. However, mental health has become a global concern, affecting individuals of all ages and backgrounds [6]. Among the common mental illnesses, anxiety disorders affect 284 million people worldwide. Depression, which often co-occurs with anxiety, affects more than 264 million people worldwide [7]. Mental health has become a critical public health concern worldwide, affecting individuals across all age groups and socioeconomic backgrounds [8]. With the fast advancement of method, AI approaches are increasingly used to address a varied range of multiple issues across different domains. The machine receives and processes new data to produce results based on facts and figs. Limiting care to symptom management rather than personalized interventions is a risk of this approach [9,10]. Elma Kerz et al. [11] developed BiLSTM models for mental health text classification are trained on human interpretable features such as readability, stylistics, lexical sophistication, cohesion, topics, syntactic complexity, sentiment and emotions. Two public benchmark datasets covering five conditions psychological stress, bipolar disorder, depression, anxiety, and ADHD were used to evaluate the models. The proposed approach enhances interpretability with LIME and AGRAD while maintaining accuracy. However, reliance on linguistic markers across datasets may reduce generalizability because variations in language and context affect consistency in early identification. Özay Ezerceli et al. [12] developed a solution for detected Natural language processing and DL method used to depression/suicidal ideation. The developed method was trained and used to three various datasets SWMH, SuicideDetection, and CEASEv2.0 using transformers and a unique model. The developed method outperformed the advanced methods in CEASEv2.0 and datasets and SuicideDetection achieved high F1 scores. However, it focused solely on Reddit data without incorporating

* Corresponding author: sujata.patil@kletech.ac.in

other social media platforms, multimodal inputs, or user characteristics such as hobbies, location, age, and gender, which restricted the scope and generalizability of the findings. Ghanadian et al. [13] developed a benchmark against advanced NLP classification models, centered specifically on the BERT family structures. A UMD real-world dataset was used, and a data-driven synthetic method was employed, informed by social factors. The advantage of this method was that it enhanced the performance and demonstrates cost-effectiveness in addressing data scarcity and ensuring diversity in data representation. However, the disadvantage was that reliance on synthetic data did not fully capture the complexity and nuances of real-world mental health language behavior, which limits generalizability. Georgios Drougkas et al. [14] employed a combined unimodal and multimodal method by various illnesses, compiling an extensive list of mental health disorder markers. The DAIC-WOZ dataset, derived from clinical interviews, was used, focusing text and audio modalities. An early fusion strategy was employed to combine the features of two constructed unimodal models before applying the machine and deep learning algorithms. The advantage of multimodal models is that they outperform unimodal models in identifying the presence of markers. However, an acoustic unimodal model requires improved feature engineering, which limits its performance. Ngumimi Karen Iyortsuun et al. [15] employed to discerned telltale symptoms from patient statements using both audio and text data. Depression detection was performed using DAIC-WOZ and EATD-Corpus datasets. A network with additive cross-modal attention with BiLSTM backbones was proposed to capture cross-modal interactions. The advantage was that the model demonstrated efficacy in detecting depression through textual and speech modalities, without preset questions. However, its application was limited to experimental settings and requires further validation for real-world therapeutic use.

The key contribution of this paper:

1. The proposed MKO-DEBiGRU method captures both contextual and semantic dependencies in mental health texts, making it suitable for multiple disorders, such as Anxiety, ADHD, Bipolar, and Depression.
2. Medical Knowledge Graph integration (MedicalKG) enhances feature representation and semantic understanding, whereas orthogonal DEBiGRU reduces redundancy and strengthens generalization.
3. SHAP-based Explainable AI (XAI) is applied to provide feature-level interpretability, allowing transparent decision-making for clinicians and users.
4. The proposed method, from preprocessing to classification, is designed for real-time mental health chat support, ensuring high accuracy, robustness, and efficiency.

1.1 Problem Statement

Accurate mental health text classification is challenging because of the limitations of traditional models in capturing both contextual and semantic dependencies in user-generated text. Unstructured and noisy inputs, including slang, abbreviations, and inconsistent phrasing, often lead to poor generalization and reduced classification accuracy. Existing methods also lack interpretability, limiting their usefulness in sensitive applications such as mental health support.

1.2 Objective

To address these challenges, this paper proposes MKO-DEBiGRU integrates DEBiGRU with a Medical Knowledge Graph and orthogonal independence. This method effectively captures bidirectional contextual features and semantic relationships, reduces redundancy, and provides explainable predictions. Enriched features were processed by the classification layer for the SWMH, UMD, and DAIC-WOZ datasets, to improve robustness, computational efficiency and accuracy.

Remainder of this paper is organized as follows: In Section II, explain about the proposed method of mental health text classification. Section III discusses the results. Section IV discusses experimental findings and results obtained. Section V talks about the conclusions.

2 Proposed Method

This paper proposes the MKO-DEBiGRU method, which integrates an orthogonal independent deep bidirectional gated recurrent unit for mental health text classification. The architecture integrates orthogonally independent BiGRU layers to extract bidirectional contextual dependencies from clinical text, thereby enhancing the feature representation and semantic understanding. To improve interpretability, a medical knowledge graph is incorporated to map model predictions to domain-specific relational knowledge. Furthermore, SHapley additive explanations (SHAP)-based explainable AI is employed to provide feature-level insights into decision making. Preprocessing involves removing stop words, punctuation, hashtags, and special characters to refine the text quality. The fused feature representations were then passed through a classification layer to accurately detect the mental health conditions. Experimental evaluation of the Anxiety, ADHD, Bipolar, and Depression datasets demonstrated that MKO-DEBiGRU achieved 99.98% accuracy, significantly outperforming existing BiLSTM models in classification performance, as shown in Figure 1.

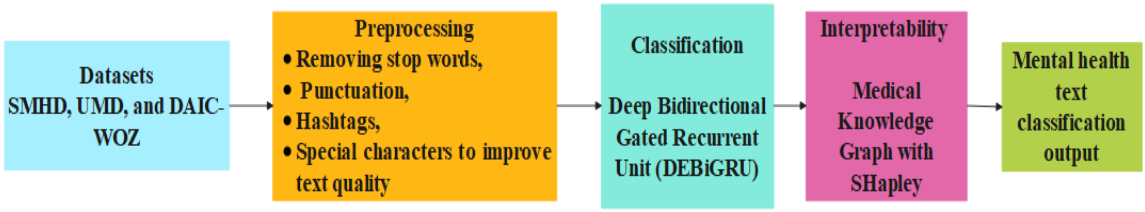

Fig. 1. Overall proposed method workflow in mental health chat support.

2.1 Dataset description

This section describes the construction and characteristics of the SWMH, UMD, and DAIC-WOZ datasets.

2.1.1 Self-Reported Mental Health Diagnoses (SMHD) Dataset

The SWMH dataset [16], also known as the SuicideWatch and Mental Health Collection, contains 54,412 posts collected using the Reddit API from depression, suicide watch, anxiety, offmychest, and bipolar.

2.1.2 University of Maryland Reddit Suicidality Dataset (UMD)

The UMD dataset [17] is another widely referenced dataset for suicidal ideation detection. TUMD dataset was formed from Reddit posts and comments expressing suicidal thoughts or behaviors by individuals. Over 100k posts and from subreddits collected comments such as Depression3 and SuicideWatch formed a dataset related to mental health and suicide prevention. Included in the data collected over a time of multiple years were the location and timing of posts, and the content of comments and posts.

2.1.3 Distress Analysis Interview Corpus-Wizard-of-Oz, and Automated interviews (DAIC-WOZ)

For this paper, DAIC-WOZ [18] dataset is utilized, is one of the four interviews in the DAIC: Automated, Wizard-of-Oz, Teleconference, and Face-to-face. Ellie, a virtual animated interviewer, carried out the DAIC-WOZ interview while being controlled in another room by a human interviewer. To support the diagnosis of mental disorders, clinical interviews including anxiety, PTSD, and depression, were contained in the dataset. The dataset contained 189 participant sessions, and the data were collected in forms including verbal, non-verbal cues, and audio.

2.2 Preprocessing

The data preprocessing pipeline was implemented for subsequent analysis, including the preprocessing steps in NLP tasks data cleaning (removing other unnecessary symbols, punctuation, hashtags, stopwords, special characters, and converting text to lowercase). To balance the labels, an oversampling technique and text tokenizing

- First, unwanted characters or symbols were eliminated in the data cleaning steps to avoid biased results. This step eliminates special characters, irrelevant information, and emojis, because they carry little significant meaning for analysis.
- Second, to balance the dataset and ensure equal representation of each class, a random oversampling technique was used. An imbalanced dataset results in a biased model, which makes this essential.
- Third, the text is tokenized into individual words, and each assigned a unique identifier as part of the preprocessing procedure. To equalize sequence lengths, zeros were padded to shorter sequences. Finally, using word embedding, the resulting tokens are transformed into numerical vectors, which allows the model to process the data more effectively and efficiently.

2.3 Classification using DEBiGRU method

$$V_{A_m} = \left[V_{A_m}, 1, \ldots, V_{A_m}.t, \ldots, V_{A_m}, T\right]^T \quad (1)$$

The presence of n-gram features at various scales in sentence phrases makes using a fixed window insufficient. It is necessary, therefore, to utilize multiple scales in parallel using ODE-BiGRU. Following the window-based sliding and stacking operation $[A_1 \ldots, A_m]$ of different scales, sentence is split into several mini-batches $U = [U_1 \ldots, U_m]$ and $[U_1' \ldots, U_m']$. Get the n-gram feature matrix U and U' as input into different ODE-GRUs, expressed as Equations (1) and (2):

$$V_t = ODE - Bi - GRU_{A_m}\left(\left[Z_{t-A_m+1,\ldots,}X_t\right], \left[Z_t, \ldots X_{t-A_m+1}\right]\right) \quad (2)$$

$V_{A_m}.t \in R^d$ denotes time series t and forward and reverse n-gram features scanned using sliding window.

Along the T-dimension of each n-gram feature matrix, each vector corresponds to a sentence.

2.4 Interpretability

This section outlines the structure and features of MedicalKG and illustrates how SHAP is applied to interpret model decisions.

2.4.1 MedicalKG

Medical KGs, constructed from medical data, contain abundant expertise that to intelligent medicine algorithms is highly helpful. Early works with relatively limited entity and relation types constructed medical KGs. A network of 4,219 human diseases based on symptoms is formed from large-scale medical bibliographic records, extracted from a single information source were most KGs. A disease–symptom graph uses rudimentary concept extraction and is constructed from 273,174 patient records. A medical knowledge network (EMKN) is formed with 6,733 nodes and 154,462 edges by EMRs for extracting medical entities. A medical KG is established with 579,094 quadruplets and 22,508 entities from single-source EMRs. From treatment data records, a traditional Chinese medicine KG is constructed and includes 59,882 entities, 17 relations, and 604,700 triples. A knowledge graph is built from multiple information sources, containing 9,230 relations, and 4,113 entities, targeted at stroke Disease.

Knowlife, constructed from various web sources contains scientific publications and online communities, is a medical knowledge graph. 8,722 symptom entities are included in the Chinese Symptom Knowledge Base (CSKB). It also has 146,631 relations between related diseases and symptoms, suffering population, stroke, and body parts. Many works have focused on embedding medical knowledge into vector representations. At the same time, to predict unknown drug interactions, various methods embed knowledge graphs. A heterogeneous graph is constructed by connecting drug databases and EMRs, containing patients, drugs, and diseases. To recommend drug based on medical diagnoses, entities are embedded simultaneously and utilized. A multi-source, large-scale medical KG, high-quality, LMKG, contains 1,225,097 relation instances and 403,784 entity instances, which are grouped into 17 relation and 13 entity types.

2.4.2 SHapley Additive exPlanations (SHAP)

SHAP is a method for interpreting predictions in a unified way, where each feature is assigned an importance value for a particular prediction. Shapley value explanation in the SHAP is denoted as linear model, and every SHAP value calculated the contribution of each feature in the model positive or negative. For all model, instant of simple linear models, SHAP offers calculation and a unique set of SHAP values for each record. In game theory, it has a solid theoretical foundation, satisfying Efficiency, Symmetry, Dummy, and Additivity, which is Shapley value. Efficiency indicates that the feature contributions must equal the difference in prediction and the average. If feature values j and k contribute equally to all possible coalitions, their contributions are the same under symmetry. Dummy gives a feature value of 0 if it never changes the predictions. Additivity sums individual feature contributions for a coalition. Finally, compared to the direct calculation of Shapley values, SHAP allows machine learning models fast computation as shows in Equation (3).

$$g(Z') = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j Z'_j \quad (3)$$

g the explanation model, Z' is coalition vector with $Z' \in 0,1^M$. A value of 1 in Z' indicates that features in new data same as original data, while 0 shows they differ. M represent maximum coalition size. For feature j in instance x, $\emptyset_j$ gives the feature attribution, which is corresponds to its Shapley value. On the prediction, feature j has a positive large impact if $\emptyset_j$ is large and positive.

3 Result and Discussion

This experiment, validated using PyTorch version 1.12.1+cu102 and Python 3.10.6 on a system with an Intel Core i7 CPU, 32 GB RAM, and an NVIDIA RTX 3090 GPU, considered the features of text and audio data.

3.1 Evaluation Metrics

The performance was evaluated for each method of mental health chat support; the definitions of these matrices are provided in Table 1.

Table 1. Evaluation metrics.

Metric	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F1 Score	$2 \times \frac{precision \times Recall}{precision + Recall}$

3.2 Performance Evaluation

The performance of proposed MKO-DEBiGRU method on three benchmark mental health datasets, namely SWMH, UMD, and DAIC-WOZ, is evaluated and compared with traditional ML and DL models, like SVM, LSTM, GRU, and BiGRU. Regarding accuracy, proposed MKO-DEBiGRU method achieved 99.98% for SWMH, 99.90% for UMD, and 99.95% for DAIC-WOZ, significantly outperforming existing models. Similarly, the proposed method demonstrates higher precision, recall, and F1-score than all baseline methods. Overall, the results indicate that the MKO-DEBiGRU model is highly effective for mental health chat support, achieving good performance across multiple datasets.

Table 2 Provides a performance comparison of MKO-DEBiGRU method with the baseline methods.

Table 2. Performance analysis of MKO-DEBiGRU.

Datasets	Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SWMH	SVM	74.8	73.5	72.9	73.2
	LSTM	78.6	77.4	76.9	77.1
	GRU	79.4	78.2	77.8	78.0
	BiGRU	81.2	80.7	80.1	80.4
	Proposed Method (MKO-DEBiGRU)	99.98	99.80	99.77	99.69
UMD	SVM	72.1	71.0	70.5	70.8
	LSTM	76.8	75.6	75.2	75.4
	GRU	80.6	76.3	75.9	76.1
	BiGRU	82.4	79.0	85.2	85.5
	Proposed Method (MKO-DEBiGRU)	99.90	98.90	99.67	98.94
DAIC-WOZ	SVM	75.5	74.1	73.8	78.2
	LSTM	79.7	78.4	78.0	79.1
	GRU	77.5	79.2	79.0	81.5
	BiGRU	79.6	88.6	88.0	88.3
	Proposed Method (MKO-DEBiGRU)	99.95	98.88	99.25	99.70

3.3 Computational Complexity

The proposed MKO-DEBiGRU method is evaluated for training and inference times and compared with the baseline models. The training time of 91 s is significantly lower than that of BERT (216 s). Similarly, the inference time of the MKO-DEBiGRU + KG-XAI model is 14 as compared to that of BERT (78 s). The results show that the MKO-DEBiGRU + KG-XAI model reduces both the training and inference times, making it highly efficient. The proposed model is computationally lightweight and is suitable for real-time mental health chat support applications. Figure 2 lists the complexity of the ML and DL models.

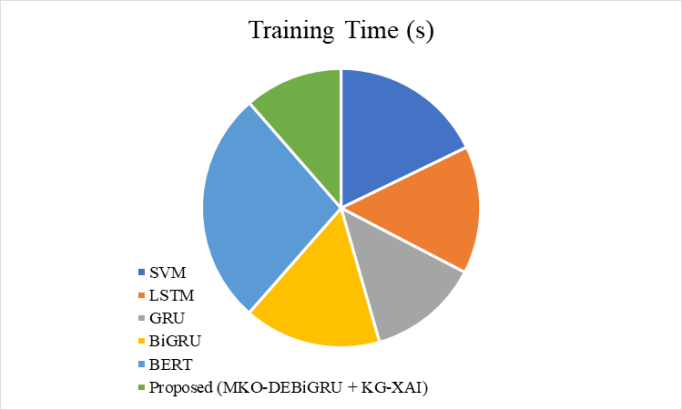

Fig. 2. Complexity of DL and ML models.

3.4 Statistical Analysis

Compared with the traditional SVM, LSTM, GRU, BiGRU, and BERT models, the proposed MKO-DEBiGRU + KG-XAI method demonstrated superior accuracy and adaptability across the SWMH, UMD, and DAIC-WOZ datasets. The improvement in the classification accuracy was statistically significant ($p < 0.05$), as validated by a paired t-test conducted over 10 independent runs. These results confirm the robustness, reliability, and effectiveness of the proposed MKO-DEBiGRU + KG-XAI method for scalable and precise mental health chat support.

3.5 Ablation Study

The ablation study in this section demonstrates that each component of the proposed MKO-DEBiGRU method enhances performance using the SWMH, UMD, and DAIC-WOZ datasets. Starting with the base BiGRU, the accuracy is comparatively lower; however, introducing the DEBiGRU architecture improved the bidirectional learning and increased the accuracy. When Medical Knowledge Graph integration is included, the method becomes more effective at capturing domain-specific semantic relations, leading to a better F1-score, precision, and recall. Finally, incorporating orthogonal independence strengthens feature learning and reduces redundancy, enabling the proposed MKO-DEBiGRU method to achieve a superior performance. Each added component contributes significantly, thereby enhancing the method to achieve high classification accuracy and reliability, as shown in Table 3.

Table 3. Ablation study of MKO-DEBiGRU.

Dataset	Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SWMH	BiGRU	80	81	77	79
	DEBiGRU	85	85	87	83
	DEBiGRU+MedicalKG	89	83	79	84
	Proposed (MKO-DEBiGRU)	99.98	99.80	99.77	99.69
UMD	BiGRU	84	89	87	84
	DEBiGRU	86	82	84	87
	DEBiGRU+MedicalKG	81	85	82	80
	Proposed (MKO-DEBiGRU)	99.90	98.90	99.67	98.94
DAIC-WOZ	BiGRU	79	81	76	77
	DEBiGRU	73	78	71	85
	DEBiGRU+MedicalKG	80	76	85	78
	Proposed (MKO-DEBiGRU)	99.95	98.88	99.25	99.70

3.6 Comparative Analysis

The proposed MKO-DEBiGRU method is evaluated on the SWMH, UMD, and DAIC-WOZ datasets and compared with the existing approaches. For the SWMH dataset, MentalRoBERTa [11] achieved 82.22% accuracy with an F1-score of 81.62, whereas BERT [12] reported an F1-score of 68. In contrast, the proposed MKO-DEBiGRU achieved 99.98% accuracy and a 99.69 F1-score, significantly outperforming previous methods. Similarly, on the UMD dataset, DistilBERT [13] obtained 77% accuracy with an F1-score of 0.75, while the proposed MKO-DEBiGRU achieved 99.90% accuracy and 98.94 F1-score, surpassing the baseline. On the DAIC-WOZ dataset, Multimodal-FCNN [14] achieved 84.59% accuracy with an 0.76 F1-score, and ACMA Net [15] reported an F1-score of 0.70. However, the proposed MKO-DEBiGRU attained 99.95% accuracy with a 99.70 F1-score, demonstrating superior performance across all datasets. These results confirm the robustness and superiority of MKO-DEBiGRU for effective mental health classification, as shown in Table 4.

Table 4. Comparative analysis with existing methods.

Datasets	Models	Accuracy (%)	F1-scores (%)
SWMH [11]	MentalRoBERTa	82.22	81.62
	Proposed (MKO-DEBiGRU)	99.98	99.69
SWMH [12]	BERT	NA	68
	Proposed (MKO-DEBiGRU)	99.98	99.69
UMD [13]	DistilBERT	0.77	0.75
	Proposed (MKO-DEBiGRU)	99.90	98.94
DAIC-WOZ [14]	Multimodal - FCNN	84.59	0.76
	Proposed (MKO-DEBiGRU)	99.95	99.70
DAIC-WOZ [15]	ACMA Net	NA	0.70
	Proposed (MKO-DEBiGRU)	99.95	99.70

4 Discussion

The experimental evaluation shows that proposed MKO-DEBiGRU method significantly outperforms the existing models on the SWMH, UMD, and DAIC-WOZ datasets. The integration of DEBiGRU and Medical Knowledge Graphs ensures the effective capture of bidirectional contextual dependencies and domain-specific semantic relationships, strengthening the feature representation for mental health classification. Orthogonal independence further reduces redundancy, enhances feature diversity, and improves the generalization across diverse datasets. The ablation study confirms that each component contributes positively to performance, with MedicalKG improving semantic understanding and orthogonal DEBiGRU achieving the optimal accuracy. Additionally, the method exhibits superior computational efficiency with reduced training and inference times compared to baseline models, making it highly suitable for real-time mental health chat support applications.

5 Conclusion

This paper proposes the MKO-DEBiGRU method, a mental health chat support system that combines DEBiGRU, Medical Knowledge Graphs, and orthogonal independence to capture both contextual and semantic features. The MKO-DEBiGRU method effectively improves feature representation, classification accuracy, and robustness. Experimental results on the SWMH, UMD, and DAIC-WOZ datasets demonstrate that MKO-DEBiGRU outperforms the existing models, achieving 99.98% accuracy for SWMH, 99.90% for UMD, and 99.95% for DAIC-WOZ. The ablation study validated the contribution of each component, whereas the computational analysis showed reduced training and inference times, confirming its efficiency. Overall, the MKO-DEBiGRU method provides a reliable, scalable, and high-performance solution for mental health chat support. Future work will focus on extending the proposed MKO-DEBiGRU model to include audio, video, and

physiological signals along with text for a more comprehensive mental health assessment.

Reference

1. Y. Liu, F. Wang, Investigating the interpretability of ChatGPT in mental health counseling: An analysis of artificial intelligence generated content differentiation. Comput. Methods Programs Biomed. **268**, 108864 (2025). https://doi.org/10.1016/j.cmpb.2025.108864
2. D. Horstkötter, M. Kanne, S. Karbouniaris, N. Lazrak, M. Bulgheroni, E. Sheltawy, L. Giani, M. La Gamba, E.R. Pujadas, M. Camacho, F. Royle, Decision-making on an AI-supported youth mental health app: A multilogue among ethicists, social scientists, AI-researchers, biomedical engineers, young experiential experts, and psychiatrists. J. Responsib. Technol. **22**, 100119 (2025). https://doi.org/10.1016/j.jrt.2025.100119
3. Z. Wang, Z. Zhang, A. Y. A. Hammadi, X. Huang, E. Damiani, C. Y. Yeun, Evolving explainable artificial intelligence for electroencephalography-based mental health classification in digital twin systems. Ad Hoc Networks. **178**, 103964 (2025). https://doi.org/10.1016/j.adhoc.2025.103964
4. A. Pesqueira, M. J. Sousa, R. Pereira, M. Schwendinger, Designing and implementing SMILE: An AI-driven platform for enhancing clinical decision-making in mental health and neurodivergence management. Comput. Struct. Biotechnol. J. **27**, 785–803 (2025). https://doi.org/10.1016/j.csbj.2025.02.022
5. L. F. Föyen, E. Zapel, M. Lekander, E. Hedman-Lagerlöf, E. Lindsäter, Artificial intelligence vs. human expert: Licensed mental health clinicians' blinded evaluation of AI-generated and expert psychological advice on quality, empathy, and perceived authorship. Internet Interv. **41**, 100841 (2025). https://doi.org/10.1016/j.invent.2025.100841
6. Z. Li, Z. Zhu, X. Gui, Y. Luo, This is human intelligence debugging artificial intelligence: Examining how people prompt GPT in seeking mental health support. Int. J. Hum. -Comput. Stud. **203**, 103555 (2025). https://doi.org/10.1016/j.ijhcs.2025.103555
7. J. Biswas, M. N. Hasan, M. S. R. Gazi, M. M. Rahman, Enhancing mental well-being: An artificial intelligence model for predicting mental disorders. Array. **26**, 100417 (2025). https://doi.org/10.1016/j.array.2025.100417
8. A. Manole, R. Cârciumaru, R. Brînzaș, F. Manole, Harnessing AI in anxiety management: A chatbot-based intervention for personalized mental health support. Information **15**, 768 (2024). https://doi.org/10.3390/info15120768
9. L. Aggarwal, V. Ranjan, A. Sharma, Real-time image processing and smart healthcare using explainable artificial intelligence (XAI). Image Vis. Comput. **161**, 105653 (2025). https://doi.org/10.1016/j.imavis.2025.105653
10. A. Babu, A. P. Joseph, Artificial intelligence in mental healthcare: Transformative potential vs. the necessity of human interaction. Front. Psychol. **15**, 1378904 (2024). https://doi.org/10.3389/fpsyg.2024.1378904
11. E. Kerz, S. Zanwar, Y. Qiao, D. Wiechmann, Toward explainable AI (XAI) for mental health detection based on language behavior. Front. Psychiatry **14**, 1219479 (2023). https://doi.org/10.3389/fpsyt.2023.1219479
12. Ö. Ezerceli, R. Dehkharghani, Mental disorder and suicidal ideation detection from social media using deep neural networks. J. Comput. Soc. Sci. **7**, 2277–2307 (2024). https://doi.org/10.1007/s42001-024-00307-1
13. H. Ghanadian, I. Nejadgholi, H. Al Osman, Socially aware synthetic data generation for suicidal ideation detection using large language models. IEEE Access **12**, 14350–14363 (2024). https://doi.org/10.1109/ACCESS.2024.3358206
14. G. Drougkas, E. M. Bakker, M. Spruit, Multimodal machine learning for language and speech markers identification in mental health. BMC Med. Inform. Decis. Mak. **24**, 354 (2024). https://doi.org/10.1186/s12911-024-02772-0
15. N. K. Iyortsuun, S. H. Kim, H. J. Yang, S. W. Kim, M. Jhon, Additive cross-modal attention network (ACMA) for depression detection based on audio and textual features. IEEE Access **12**, 20479–20489 (2024). https://doi.org/10.1109/ACCESS.2024.3362233
16. SWMH dataset Link: https://www.kaggle.com/datasets/souvikahmed071/social-media-and-mental-health (Accessed on 30.08.2025)
17. UMD dataset Link: https://www.kaggle.com/datasets/thedevastator/c-ssrs-labeled-suicidality-in-500-anonymized-red (Accessed on 01.09.2025)
18. DAIC-WOZ dataset Link: https://www.kaggle.com/datasets/abdelrahmanahmed3/daic-woz-audio9 (Accessed on 01.09.2025)