

Sentiment Description-Based Multi-Level Depression State Prediction and Severity Regression: A Holistic Mental Health Clues Fused with Lexicon-Driven Features

Soojin Lee, Sukhan Lee

Submitted to: JMIR Mental Health
on: October 12, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 33

 Multimedia Appendixes..... 34

 Multimedia Appendix 1..... 34

 Multimedia Appendix 2..... 34

Sentiment Description-Based Multi-Level Depression State Prediction and Severity Regression: A Holistic Mental Health Clues Fused with Lexicon-Driven Features

Soojin Lee¹; Sukhan Lee² PhD

¹ Sungkyunkwan University Suwon KR

² Sungkyunkwan University Suwon-si, Gyeonggi-do KR

Corresponding Author:

Sukhan Lee PhD

Sungkyunkwan University
2066, Seobu-ro, Jangan-gu
Suwon-si, Gyeonggi-do
KR

Abstract

Background: The increasing prevalence of mental health problems among populations prompts the necessity of a mental health self-monitoring platform for the general public as an intermediary to professional treatment, which enables self-diagnosis of mental disorders based on natural lifelog data from daily living environments.

Objective: In this study, we proposed a novel approach to self-predicting depression states with their severity levels based on naturally collected emotional narrative text, referred to here as “sentiment description.”

Methods: Unlike conventional text-based approaches, which are mainly lexicon-driven, we extracted holistic mental health clues from the contextual connotations embedded in sentiment descriptions. By integrating them with lexicon-driven features, we aimed to enhance the reliability and explainability of depression diagnosis, supported by rich evidence from various angles. In particular, the holistic mental health clues extracted are designed to have a direct link to the standard mental health indices for diagnosing depression, the level of depression-indicative sentiments associated with chosen aspects, and the personal psychological tendencies affecting the sentiment description-based self-diagnosis network (SD2Net). In SD2Net, 1) holistic clues are extracted based on a large linguistic foundation model, GPT, with optimally designed prompts, 2) the holistic clues and raw texts are transformed into the connotation-driven and lexicon-driven feature embeddings are fused into multi-level depression state prediction, together two learned embeddings are fused into multi-level depression state prediction, together with the regression of their severity-levels, based on a hierarchical decision-making framework. For the experiment, we used both a benchmark dataset and a dataset custom-collected from real subjects on a long-term basis. Notably, we annotated the custom-collected data by fusing multiple self-questionnaires and psychiatrists’ inputs to ensure the trustworthiness of ground truth labels.

Results: The experiments demonstrate the state-of-the-art performance of the proposed approach for holistic clue-reinforced multi-level depression state prediction, achieving an accuracy of approximately 80% using both a custom-collected dataset and the DepSign benchmark dataset. The ablation study confirms the effectiveness of the proposed holistic clue model, which yields an average improvement of over 10% in accuracy compared to models without it.

Conclusions: The proposed approach, SD2Net, shows the effectiveness of incorporating holistic mental health clues from sentiment descriptions for multi-level depression state prediction, highlighting its potential as a reliable self-monitoring tool for the general public.

(JMIR Preprints 12/10/2025:85715)

DOI: <https://doi.org/10.2196/preprints.85715>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://www.jmir.org/preprint/85715>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/preprint/85715>



Original Manuscript

Original Paper

Sentiment Description-Based Multi-Level Depression State Prediction and Severity Regression: A Holistic Mental Health Clues Fused with Lexicon-Driven Features

Abstract

Background: The increasing prevalence of mental health problems among populations prompts the necessity of a mental health self-monitoring platform for the general public as an intermediary to professional treatment, which enables self-diagnosis of mental disorders based on natural lifelog data from daily living environments.

Objective: In this study, we proposed a novel approach to self-predicting depression states with their severity levels based on naturally collected emotional narrative text, referred to here as “sentiment description.”

Methods: Unlike conventional text-based approaches, which are mainly lexicon-driven, we extracted holistic mental health clues from the contextual connotations embedded in sentiment descriptions. By integrating them with lexicon-driven features, we aimed to enhance the reliability and explainability of depression diagnosis, supported by rich evidence from various angles. In particular, the holistic mental health clues extracted are designed to have a direct link to the standard mental health indices for diagnosing depression, the level of depression-indicative sentiments associated with chosen aspects, and the personal psychological tendencies affecting the sentiment description-based self-diagnosis network (SD2Net). In SD2Net, 1) holistic clues are extracted based on a large linguistic foundation model, GPT, with optimally designed prompts, 2) the holistic clues and raw texts are transformed into the connotation-driven and lexicon-driven feature embeddings are fused into multi-level depression state prediction, together two learned embeddings are fused into multi-level depression state prediction, together with the regression of their severity-levels, based on a hierarchical decision-making framework. For the experiment, we used both a benchmark dataset and a dataset custom-collected from real subjects on a long-term basis. Notably, we annotated the custom-collected data by fusing multiple self-questionnaires and psychiatrists’ inputs to ensure the trustworthiness of ground truth labels.

Results: The experiments demonstrate the state-of-the-art performance of the proposed approach for holistic clue-reinforced multi-level depression state prediction, achieving an accuracy of approximately 80% using both a custom-collected dataset and the DepSign benchmark dataset. The ablation study confirms the effectiveness of the proposed holistic clue model, which yields an average improvement of over 10% in accuracy compared to models without it.

Conclusions: The proposed approach, SD2Net, shows the effectiveness of incorporating holistic mental health clues from sentiment descriptions for multi-level depression state prediction, highlighting its potential as a reliable self-monitoring tool for the general public.

Keywords: Mental Health; Depression; Self-diagnosis; Self-monitoring; Lifelog Data; Natural Language Processing; Holistic Mental Health Clues

Introduction

According to the World Health Organization (WHO), approximately one in eight people suffers from some form of mental disorder worldwide. Recently, the United Nations (UN) reported a remarkable 25-27% increase, in particular, in depression and anxiety disorders [1]. In South Korea, the Ministry

of Health and Welfare has reported that the prevalence of mental disorders among the total population has reached 27.8% in 2021 [2]. This has given rise to a demand for a new and innovative approach to monitoring, diagnosing, and treating mental disorders of the general population [3] to promote more effective and timely mental health services and management for the public. Specifically, this prompts a need for a shift from traditional in-person psychiatric care to a new paradigm based on remote psychiatric diagnosis and treatment mediated by self-monitoring platforms based on natural lifelog data collected in daily living environments [4].

Two distinct perspectives drive the need for a non-face-to-face self-monitoring platform that collects natural lifelog data in a daily living environment. First, the conventional self-reporting method, e.g., based on psychiatrists' interviews and diagnostic self-questionnaires, relies on subjective reports from patients, which are based on their memory and are therefore susceptible to memory distortion, underreporting, and concealment, thereby limiting the quality of the data [5]. Second, the utilization of self-monitoring platforms in daily living environments facilitates the more efficacious establishment of public mental health services and management, providing self-diagnosis and self-treatment tools for the general population, including specific vulnerable groups. Notably, non-face-to-face-based mental health services have evolved beyond simply providing information requested by users or referring them to traditional mental health services [6]. They now offer far advanced services that utilize digital indicators and therapies from mobile and smart devices [7]. In particular, digital indicators extracted from natural lifelog data have the potential to enhance the objectivity of clinical evaluations by compensating for the aforementioned limitations of traditional self-reporting methods [5], as well as facilitating the timely detection of mental disorders. Accordingly, recent non-face-to-face mental health services are evolving beyond a simple remote provision of requested information into a strong emphasis on real-time self-monitoring of mental disorders using digital indicators from natural lifelog data collected by smart devices as key diagnostic elements [8]. Having said that, however, despite advances in self-monitoring platforms based on natural lifelog data for real-time mental health services and management, a comprehensive scientific framework on which their validity and effectiveness in psychiatric diagnosis and treatment should be grounded remains to be further established.

As for the real-time remote mental health monitoring based on lifelog data, recently, the self-prediction of mental depression levels has drawn particular attention. In particular, lifelog data that captures individuals' expressions of emotions, thoughts, and experiences in their daily lives without restriction, which we referred to here as "sentiment description," can provide a rich source of information on mental depression states among various digital phenotypes [9]. For instance, compared to traditional self-report questionnaires, which rely on predetermined questions, Sentiment descriptions can provide more diverse yet less corrupted expressions of one's emotional and mental states. This makes them a particularly effective digital phenotype for predicting, analyzing, and interpreting an individual's level of depression.

Related Work

Among various types of lifelog data, deep learning-based studies on depression stage prediction using textual information can broadly be categorized into two approaches, depending on the prediction objective. The first approach involves binary classification, distinguishing between depressed and non-depressed states using text data [10, 11, 12]. The second approach focuses on predicting finer-grained levels of depression [13, 14, 15, 16, 17]. In these studies, textual data collected through hashtags or depression-related forums are annotated by trained experts using the criteria for Major Depressive Disorder (MDD) defined in the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders) [18] or through user-submitted self-report questionnaires. Although most current research focuses on binary classification tasks due to the large data requirements of

deep learning, some studies have attempted to predict multi-level depression. This GT labeling arises from the difficulty of constructing medically verified GT labels for large-scale, unstructured datasets by mental health professionals. The detailed contents of representative studies on lifelog text-based depression prediction are summarized in Table 1.

Table 1. Summary of representative studies on text-based depression prediction (binary vs multi-level approaches).

Approach	Authors	Dataset (Source)	Model	Performance	Purpose / Contribution
Binary	Amanat et al. [10]	Twitter (Crawled)	LSTM [19], RNN	Acc. 99.66%	Feasibility of text-based depression detection. One-hot encoding + PCA for feature extraction.
	Bokolo et al. [11]	Sentiment140 Twitter [24] (Crawled)	DistilBERT [20], SqueezeBERT [21], DeBERTa, [22], RoBERTa [23]	Acc. 97%	Compared transformer models on large-scale Twitter data. Analyzed correlations with linguistic patterns.
	Shin et al. [12]	91 participants' diaries (App)	GPT-3.5, GPT-4	Acc. 90.2%,	Utilized LLMs for PHQ-9 [25] and BSSI [26] depression prediction based on diary-style text data.
Multi-Level	Milintsevich et al. [13]	DAIC-WOZ [28] (Benchmark)	Transformer + LSTM	MAE 3.78, F1-Score 73.9%	Predicted PHQ-8 [27] item scores from interview data resembling PHQ-8 flow, providing initial diagnostic-level results.
	Muñoz et al. [14]	DsD [29], DepSign [30] (Benchmark)	Lexicon + Embedding + ML & Transformer	F1-Score: 70% (DsD), 69% (DepSign)	Improved performance by combining lexicon-based and embedding-based features.
	Kabir et al. [15]	Twitter (Crawled)	BERT [31], DistilBERT	Acc. 72.5%	Enhanced performance using word embeddings, sentiment analysis, and sentence-level feature extraction.
	Poświata et al. [16]	LT-EDI-ACL2022 DepSign (Benchmark)	RoBERTa Ensemble	F1-Score 58.3%	Enhanced performance by fine-tuning with a depression-specific corpus [32].
	Kim et al. [17]	MHB [33], CAMS [34], HelaDepDet [35], RMHD [36], DepressionE mo [37] (Benchmark)	Zero-Shot LLM (Prompt Engineering) LLM-Derived Embedding (Summary) + ML	Binary: Acc. 96% Multi-State: Acc. 58%	Overcame limitations of traditional text embeddings (lack of clinically relevant emotional cues) using LLMs. Summarized raw text with GPT from a mental health perspective, then applied data-driven methods.

Conventional approaches to predicting mentally depressive states based on text data primarily rely on lexicon-associated features extracted through deep processing of sentence sequences. In

particular, a transformer model [10–17] is preferred for deep processing to leverage its capability to contextually associate individual lexicons within sentences, which enables the extraction of context-reinforced, depression-related features. In other words, a transformer model is capable of learning depression-related contextual features in the latent space, such that the contextual features thus formed in the latent space are more effective for discriminating out the mentally depressive states embedded in the sentences. Although they are shown to be effective, we considered that the conventional lexicon-based approaches are somewhat limited in terms of exploiting a higher level of contextual connotations, sentiments, and personal tendencies embedded in text data, which can only be extracted by holistic abstraction and assessment of text data into meaningful clues for predicting mentally depressive states. Specifically, we summarized the following limitations of conventional lexicon-based approaches in depression self-diagnosis: 1) the lexicon-based features extracted may have a limited relevance to the medical criteria used by psychiatrists for clinical diagnosis of depression; 2) the text data used for depression prediction are collected mainly from social media by crawling with their subjects covering widely different aspects, making it highly vulnerable to the issue of data dependency; 3) the lexicon-based approaches are limited in incorporating a higher level of contextual and circumstantial evidences embedded in the text that are relevant to mentally depressive states; and 4) the lexicon-based approaches offer a limited capacity in explainability that can promote a richer clinical value beyond the conventional visualization of lexicon heat maps according to their levels of contribution to prediction.

Besides the above limitations, most conventional approaches to date using text data for depression diagnosis have focused on classifying between non-depression and depression, rather than predicting the level of depression states, as often practiced in psychiatric clinics. Furthermore, their prediction results are considered unreliable, primarily due to the large data imbalance in the training dataset and the lack of evidence regarding the medical validity of their results. For example, most previous studies rely on user-generated posts from online social media platforms like Twitter and Reddit, which makes it challenging to obtain high-quality data that is well-balanced across different levels of depression states and accurately annotated to guide the GT states. In most cases, annotations rely on either trained specialists or self-posted diagnoses without a direct involvement of professional psychiatric assessment. Note that, as introduced, an automatic prediction of depression states based on text data is targeted for use in a depression self-monitoring platform that automatically links the public to psychiatrists as an intermediary toward real-time monitoring and treatment of public mental health. Therefore, despite existing challenges, there is a need to enhance reliability and explainability beyond conventional lexicon-based approaches for text-based depression state predictions, as they play a critical role as an intermediary in a mental health self-monitoring platform for the public.

Contribution

The main contribution of this study lies in incorporating a higher level of contextual connotations of the text into conventional lexicon-based features as additional yet holistic clues for depression diagnosis. In particular, we aimed to extract clues directly related to the standard medical criteria for diagnosing depression, the depression-indicative sentiment levels associated with the chosen aspects, and the personal psychological tendencies embedded in the text. We aimed at enhancing the reliability and explainability of text-based depression diagnosis beyond what conventional lexicon-based approaches can offer by supplementing them with contextual connotation-based holistic clues, in analogy to the process in which psychiatrists reach a diagnostic decision by collecting holistic evidence, both direct and circumstantial, based, e.g., on standard questionnaires, personal interviews, and observations of patient's behavior and appearance. Ultimately, we envisioned that the enhanced reliability and explainability contributed by the proposed approach would help enable text-based depression self-diagnosis to play a key role for mental health self-monitoring platforms as an intermediary to professional psychiatric treatment.

In addition to the conceptual expansion of the text-based depression self-diagnosis framework, we proposed several distinct approaches for implementing the conceptually expanded framework. First, we extracted numerous connotation-driven holistic clues from the text that represent the levels associated with standard medical criteria, depression-indicative sentiments, and personal psychological tendencies, using GPT with optimized prompts. For convenience, the extracted connotation-driven holistic clues are referred to here as “holistic mental health clue (HMHC) indices.” We transformed the extracted HMHC indices into their feature embedding, “connotation-driven embedding”, by a text-based visual transformer (Text-ViT), so that the holistic clue-driven embedding can be concatenated with the lexicon-driven feature embedding generated by BERT for the final diagnostic decision. Second, we extended the conventional binary diagnostic decision, which is either “normal” or “abnormal”, into a multi-level decision that includes “normal”, “mild”, and “moderate/severe”. In particular, to ensure accuracy in multi-level diagnostic decisions under possible data imbalance, we introduced a hierarchical decision-making framework in which abnormal cases are further classified into mild and moderate/severe subclasses. Furthermore, we output the diagnostic decision not only in terms of classification but also in terms of regression, indicating the severity level associated with predicted depression states for a more explainable undertaking in self-monitoring. Third, for the experiment, we custom-collected a dataset from real subjects on a long-term basis, in addition to a benchmark dataset, which is annotated by fusing multiple self-questionnaires with a psychiatrist’s inputs for the trustworthiness of GT labels [41].

Methods

Overall System

As described, we aimed to extract as many contextual and circumstantial clues as possible from the text of sentiment descriptions to incorporate the connotation-driven holistic feature embedding into the lexicon-driven atomistic feature embedding for the final diagnostic decision. Specifically, the holistic clues we extracted from the text are represented as the following indices, referred to here as “Holistic Mental Health Clue (HMHC)” indices: 1) the standard mental health indices used in psychiatric practices for diagnosing mental health states, 2) the aspect-sentiment indices representing the level of sentiment expressed in the text, where the aspects are custom-selected to organize the widely diversified written subjects into their simplified high-level categories, 3) the depression-related experience indices as an indicator of personal experience in depression diagnosis and treatment, and 4) individual psychological traits indices in terms of positive and negative attitudes. A large linguistic foundation model (LLM), such as GPT, with optimally designed prompts, is used to extract the above HMHC indices. For implementation, we proposed a deep network architecture, referred to here as the sentiment description-based self-diagnosis network (SD2Net), that learns the HMHC-driven feature embedding from the extracted HMHC indices based on Text-ViT (Vision Transformer [38]) and the lexicon-driven feature embedding based on BERT, such that both feature embeddings are fused into a more reliable and explainable diagnostic decision under a hierarchical decision-making framework. In the following subsections, we present the details of the methodologies implemented for the individual components of SD2Net.

SD2Net Architecture

SD2Net is a predictive model that quantifies an individual’s sentiment description into HMHC feature indices, transforms HMHC feature indices into connotation-driven feature embedding, extracts lexicon-driven feature embedding from the raw text, and integrates both embeddings to perform fine-grained prediction of depression states by regression. As illustrated in Figure 1, the system consists of the following core modules: 1) HMHC feature index extraction module leveraged by GPT-3.5 [39]. The HMHC feature index module transforms a general sentiment description into

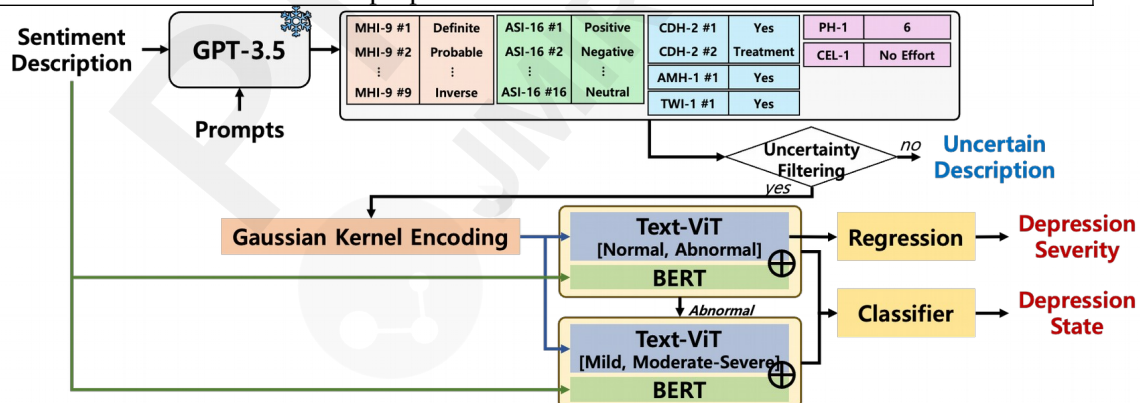
seven types of high-level, mental health-oriented features. They are “Mental Health Index-9 (MHI-9), Aspect Sentiment Index-16 (ASI-16), Clinical Diagnosis History-2 (CDH-2), Antidepressant Medication History-1 (AMH-1), Treatment Willingness Index-1 (TWI-1), Perceived Hopefulness (PH-1), and Coping Effort Level-1 (CEL-1)”. These indices provide holistic indicators tailored to psychological and psychiatric assessment.

2) Depression state prediction module fusing the HMHC-driven feature embedding and the lexicon-driven feature embedding. The HMHC-driven feature embedding is learned by Text-ViT, using HMHC feature indices obtained as input via GPT. In contrast, the lexicon-driven feature embedding is learned by BERT with the raw sentiment descriptions as its input. Note that, before inputting the HMHC indices from GPT, SD2Net counts “don’t know” on the extracted HMHC feature indices to uncertainty-check if meaningful holistic clues are available in sentiment descriptions. The HMHC feature indices passed the uncertainty check and then underwent Gaussian kernel encoding for enhanced expressiveness before being input to Text-ViT.

3) Depression prediction in SD2Net is performed hierarchically, with the normal-abnormal classification in the first layer and the mild-moderate/severe classification of the data identified as abnormal in the second layer. To this end, the same combination of Text-ViT and BERT architecture is used for both layers. For each layer, the HMHC-driven and lexicon-driven feature embedding outputs from the Text-ViT and BERT models, respectively, are concatenated and passed into a multi-layer perceptron (MLP) to classify the depression state.

4) Depression severity-level prediction based on a regression network module. In addition to predicting the depression state into non-depressed, mild, moderate-severe classes, a regression model is attached to estimate the severity level associated with predicted depression classes. This module utilizes the first-layer Text-ViT embedding to learn its position in the annotated GT space formed by multiple self-questionnaires, CES-D [40], and PHQ-9, as well as psychiatrists’ input, thereby providing a more detailed assessment of severity with explainability.

Figure 1. Overall architecture of the proposed SD2Net.



GPT-based HMHC Index Extraction

Lifelog data typically consists of free-form descriptions in which individuals express their thoughts and emotions across a range of topics without constraints in format or structure. The responses given are based on each individual’s personal feelings about their experiences. Given the nature of such data, where follow-up clarification or consistent questioning is not possible, we propose a novel approach that extracts mental health-related features.

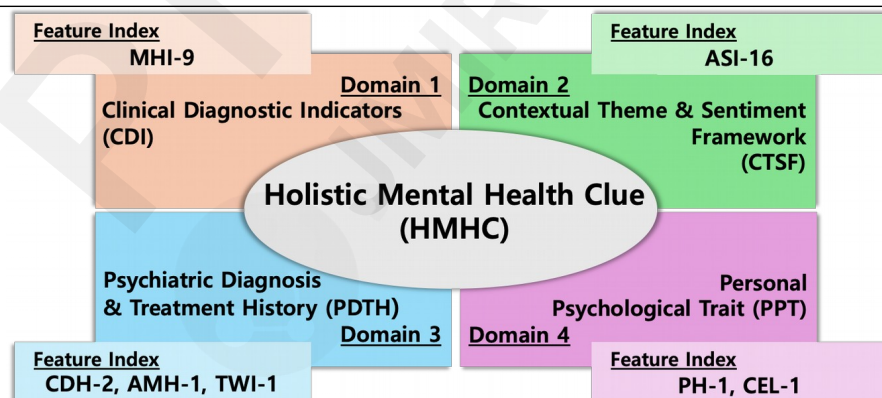
Conventional approaches have primarily employed methods such as calculating the VAD (Valence,

Arousal, Dominance) scores or polarity scores derived from raw text data. These studies have also employed the extraction of feature embeddings based on the sentiment of words used in sentences, as well as the extraction of feature embeddings specialized for specific topics based on a lexicon. However, these extraction methods are more focused on mood information rather than being specialized for mental health, which is a disadvantage, as they do not act as useful features for predicting states of depression. As a result, they are limited in their ability to support detailed depression state predictions and often fail to extract complex, high-level affective cues present in free-form text.

In the proposed approach, we defined four domains associated with mental health, and based on their characteristics, we further established seven sub-feature indices. Figure 2 provides a schematic representation of the four domains together with the feature indices defined for each. The seven categories of mental health feature indices encompass both clinically utilized depression-related indicators and broader dimensions such as general life themes and personal attitudes. The four domains are: Clinical Diagnostic Indicator (CDI), Contextual Theme & Sentiment Framework (CTSF), Psychiatric Diagnosis & Treatment History (PDTH), and Personal Psychological Trait (PPT).

The CDI domain concerns indicators commonly employed in clinical settings for the diagnosis of depression, with the Mental Health Index-9 (MH-9) defined as its sub-feature index. CTSF captures general themes that appear in lifelogs along with the sentiments expressed toward them, represented by the Aspect Sentiment Index-16 (ASI-16). PDTH reflects individual experiences related to mental health and includes three sub-feature indices: Clinical Diagnosis History-2 (CDH-2), Antidepressant Medication History-1 (AMH-1), and Treatment Willingness Index-1 (TWI-1). Lastly, PPT refers to personal psychological traits, with Perceived Hopefulness-1 (PH-1) and Coping Effort Level-1 (CEL-1) defined as its sub-feature indices. The details of the HMHC feature indices described in this study are summarized in Multimedia Appendix 1.

Figure 2. Schematic representation of the HMHC four domains with the feature indices defined for each.



The proposed seven HMHC feature indices are robust in distinguishing not only between depressed and non-depressed individuals but also between subgroups within the depressed population (mild, moderate, severe). In contrast to prevailing lexicon-based data-driven feature embedding methods, this approach is tailored for mental health applications and yields holistic clues that are both interpretable and actionable. This capability facilitates practical diagnostic assessments of depression and functions as an intermediary between patients and healthcare providers on self-monitoring platforms. However, it has been observed that GPT does not consistently generate precise responses; instead, it offers reactions that, while sincere, may deviate in content or contain inaccuracies.

Consequently, GPT's responses are characterized by the presence of opinions and hallucinations that lack a factual foundation. Therefore, it is imperative to construct GPT prompts optimally to facilitate the extraction of pertinent information while minimizing extraneous noise. We tested multiple prompt versions to determine the most effective prompting strategy for extracting reliable index values.

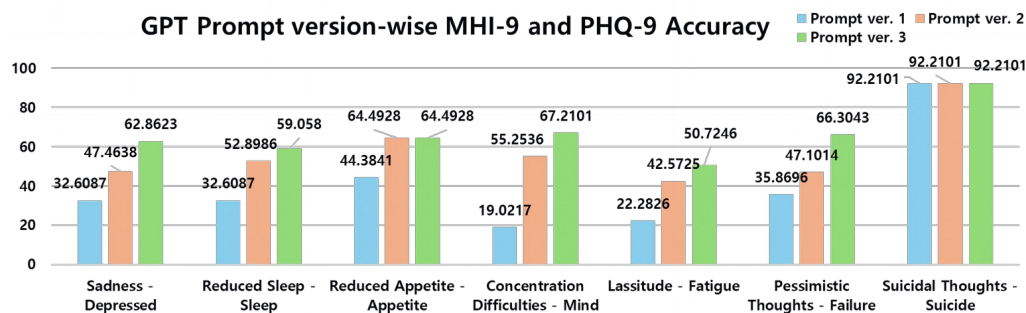
A custom-collected dataset of Korean university students was used to validate the GPT prompt. It includes a PHQ-9 self-report questionnaire responses and sentiment description texts synchronized. This allowed us to assess the validity of GPT-generated MHI-9 values by analyzing correlations with actual PHQ-9 item scores. Table 2 summarizes the correspondence between MHI-9 and PHQ-9 items, excluding "Inner Tension" and "Inability to Feel," for which no direct PHQ-9 counterparts exist.

Table 2. PHQ-9 questions correspond to MHI-9 sub-items.

MHI-9 Item Number	MHI-9 Item Content	PHQ-9 Item Number	PHQ-9 Item Content
1	Sadness	1	Depressed
2	Inner tension	-	-
3	Reduced sleep	3	Sleep
4	Reduced appetite	4	Appetite
5	Concentration difficulties	8	Mind
6	Lassitude	6	Fatigue
7	Inability to feel	-	-
8	Pessimistic thoughts	7	Failure
9	Suicidal thoughts	9	Suicide

The accuracy of three versions of GPT prompts was evaluated using the values of seven items corresponding to MHI-9 and PHQ-9, respectively. The initial prompt serves as a method for extracting all nine MHI-9 sub-items with a single prompt. However, when all nine sub-items were presented simultaneously, the accuracy of the answers to the sub-items located in the middle of the prompt sequence decreased. As illustrated by the blue graph in Figure 3, the lower accuracy of each sub-item for the initial version of the GPT prompt is depicted. To address this challenge, the second prompt inquired about each of the nine sub-items independently. As a result, as shown in the orange graph in Figure 3, the overall accuracy increased compared to the first version. Additionally, the two items (Concentration Difficulties, Lassitude) with lower accuracy also achieved accuracy levels similar to those of the other items. However, a subsequent analysis of the values extracted through GPT revealed instances where content not directly related to depressive symptoms (e.g., physiological effects of substances or other medical conditions) was considered. Additionally, there were instances where no accurate mention was made, yet a judgment was made based on speculation. To overcome this, the third prompt was designed to focus on parts directly related to depressive symptoms when analyzing the given sentence, and to use only the content accurately described in the given sentence to prevent speculation or hallucination. As demonstrated in the green graph of Figure 3, the accuracy for all sub-items surpassed 50%, and the overall accuracy attained 66.12%.

Figure 3. Sub-item accuracy for MHI-9 and PHQ-9 pairs across three GPT prompt versions.



Through extensive prompt engineering, we found that the most reliable outputs were obtained by asking specific sub-item questions while including explicit instructions to suppress GPT's opinion or hallucination. Although these prompts are effective in reducing error, we note that no single prompt can be universally optimal for all input cases. The finalized main format of the prompts for extracting values for the HMHC feature index is as follows:

You are a clinical evaluator analyzing text related to mental health. Based on the user's written input, assess the level of the 'sub-item name'.

△ Important Condition: Disregard symptoms or explanations that could be attributed solely to the physiological effects of a substance (e.g., drug use, medication) or another medical condition. Please do so based solely on what is clearly stated in the sentence, completely eliminating any guesswork from the context.

📊 Output Instructions: Score the severity of distress/impairment on an α to β scale, using the following guideline:

α : α 's definition ...

β : β 's definition.

Respond ONLY with the score (a single digit $\alpha \sim \beta$): 'RAW Text'

Multi-Level Depression State Prediction Based on HMHC-driven and Lexicon-Driven Feature Embeddings

Preprocessing of HMHC Feature Index Values via Gaussian Kernel Encoding

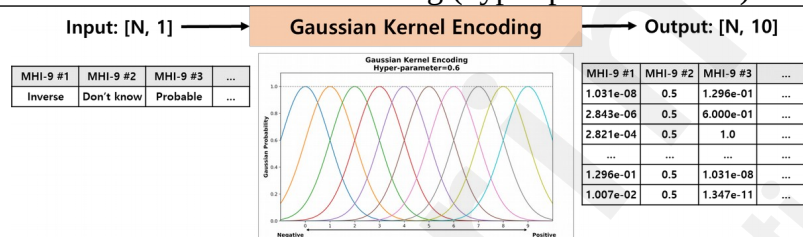
Before feeding the extracted HMHC feature index values into the deep learning modules for depression state and severity prediction, we apply an uncertainty filtering process to the index values of feature indices. Each of the seven HMHC feature indices includes sub-items that may be labeled as "Don't know" when GPT is unable to assign a value confidently. There are up to 26 such sub-items across all indices, and our analysis revealed a significant decline in model prediction accuracy when more than approximately 69% (18 or more) of the items are marked as "Don't know." Based on this observation, we set a threshold. If an input sample contains 18 or more "Don't know" labels, the system abstains from making a prediction and outputs "Unable to Determine" instead. This filtering method extends conventional input quality filtering, which is typically based on sentence length, to a higher-level semantic filtering approach.

After filtering, the selected input samples undergo Gaussian kernel encoding to transform discrete HMHC index values into probabilistic distributions. Lifelog data in SD2Net primarily capture emotional content, and most HMHC index values represent continuous rather than categorical concepts. Therefore, it is more effective to express them using Gaussian probability rather than one-hot encoding. Hence, by incorporating Gaussian kernel encoding, we can account for the possibility that other classes may also exist at the same index value, enabling a more precise prediction of depressive states and their severity based on as much information as possible. Two of the seven

HMHC indices—PH-1 and CEL-1—are already defined as continuous values. Thus, these two are represented using one-hot encoding without the Gaussian transformation.

Figure 4 illustrates the process of converting index values into probabilistic distributions using Gaussian kernel encoding. Each sub-item of the HMHC feature indices is passed through a Gaussian kernel, resulting in a 10-dimensional probability vector. After constructing Gaussian probability distributions for each value class, we set the intersection point between each probability distribution as a hyperparameter to generate a Gaussian kernel. It is evident that the magnitude of the hyperparameter value directly correlates with the degree of overlap between probability distributions. Furthermore, the positions of each class distribution are ordered based on the semantic relationships between class values, such that opposing classes (e.g., “Heavily Negative” vs. “Heavily Positive”) are positioned farther apart in the probability space.

Figure 4. The process of Gaussian Kernel Encoding (hyper-parameter: 0.6)



Depression State Prediction by Fusing Text-ViT and BERT Feature Embeddings

The HMHC feature indices, which are continuously and probabilistically transformed through Gaussian kernel encoding, are input into the hierarchical Text-ViT to extract feature embeddings corresponding to each HMHC feature index. The hierarchical Text-ViT is a modified version of the Vision Transformer (ViT), initially designed for image domains, adapted in this study for use in the text domain. Unlike standard ViT, where image patches are used as input tokens, this model utilizes Gaussian probability distributions of the HMHC feature indices as input patches. The input dimension is defined as 10, and feature indices with fewer than 10 classes are zero-padded in the remaining dimensions. As illustrated in Figure 5 (a), the proposed SD2Net leverages not only the GPT-based HMHC feature indices but also incorporates BERT-based data-driven feature embeddings to predict depression states. To extract the data-driven feature embeddings, raw text data is input into BERT, and embeddings are obtained from the final encoder layer. In this process, BERT is initialized with pre-trained weights from a sentiment classification task. The lower eight encoder layers are frozen to retain the pre-trained weights, while the top 4 encoder layers are fine-tuned through transfer learning based on the depression state prediction task in conjunction with the Text-ViT. Finally, the embeddings of HMHC feature indices extracted from the transformer layers of Text-ViT are concatenated with those from BERT, and the combined representation is passed through an MLP to predict the current state of depression.

Enhancing Prediction Performance via HMHC-based Feature Clustering and Hierarchical Decision-Making

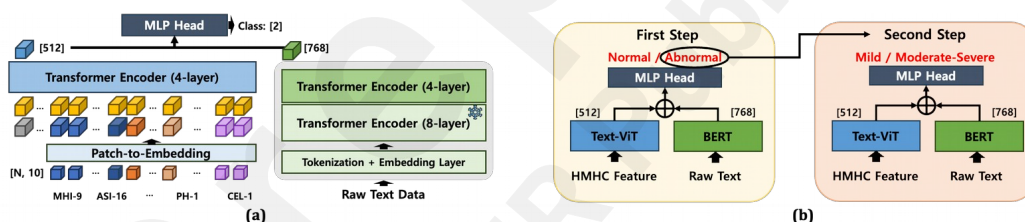
Traditional studies have commonly used user-generated posts from online platforms such as Twitter or Reddit. However, these datasets are composed of texts from an unspecified general population, which poses a limitation in obtaining balanced and high-quality data across different states of depression. To address the issue of data imbalance structurally, we propose two strategies.

The first strategy clusters the dataset into four subgroups based on responses to sub-item 1 of CDH-2

and AMH-1. The subgroups are defined as follows: (1) unknown status of depression diagnosis and unknown experience of antidepressant medication (No/No group); (2) unknown status of depression diagnosis but with experience of antidepressant medication (No/Yes group); (3) with a history of depression diagnosis but unknown experience of antidepressant medication (Yes/No group); and (4) with both a history of depression diagnosis and experience of antidepressant medication (Yes/Yes group). While each subgroup utilizes the same deep learning model architecture, the models are trained independently and maintain separate sets of weights. Note that for the Yes/Yes group, to prevent BERT from over-attending to already standard features in this group during feature extraction, the raw text inputs for BERT are modified by removing sentences that include medication-related terms using a predefined lexicon. However, the GPT-based HMHC feature values are extracted from the original unmodified raw texts.

The second strategy adopts a hierarchical (two-step) model structure: in the first step, the model performs binary classification to distinguish between “Normal” and “Abnormal” depression states. In the second step, the model predicts specific depression states only for instances classified as “Abnormal” in the previous step. By decomposing the overall task into smaller and simpler subtasks, rather than forcing a single model to perform a complex multi-class prediction at once, this hierarchical modeling strategy maximizes performance through localized learning. Figure 5 (a) illustrates the architecture of a single Text-ViT and BERT integrated module, while Figure 5 (b) presents a flow diagram of the group-wise and hierarchical prediction process using the integrated Text-ViT and BERT module.

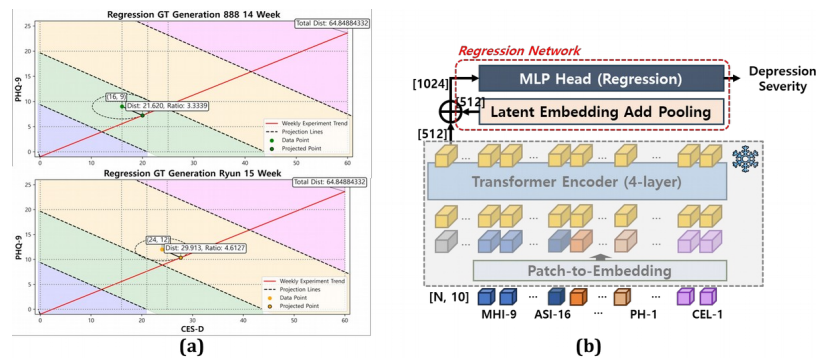
Figure 5. Depression prediction structure: (a) Single Text-ViT network structure. (b) Model flowchart for stepwise prediction.



Depression Severity Prediction Based on Regression Network

Depression state prediction is formulated as a three-class classification task consisting of Normal, Mild, and Moderate. However, for the developed SD2Net to serve effectively as an intermediary within a mental health self-monitoring platform, it must provide not only categorical predictions of depression states but also explanatory or analytical insights into the current condition. To this end, a regression network is attached to the final transformer encoder layer of the first-step Text-ViT module to estimate the severity of the current depression state. Figure 6 (a) illustrates the process for generating the target severity value. For a fused CESD-PHQ9 point (represented as a yellow point with a black outline) located on the regression trend line (in red) derived from CES-D and PHQ-9 scores, the Euclidean distance from the origin is calculated. It is essential to note that the proposed depression severity prediction model is based on a dataset collected explicitly for this study from Korean university students, utilizing the CES-D and PHQ-9 as self-report questionnaires.

Figure 6. Visualization of the depression severity prediction process: (a) Example of GT value generation, (b) Architecture of the regression network used for predicting depression severity.



As illustrated in Figure 6 (b), the regression network for predicting the severity of depression is implemented as an MLP. The input to the MLP is constructed by performing additive pooling on the latent embeddings output from the final transformer encoder layer of the first-step Text-ViT, followed by concatenation with the class embedding. During training of the regression network, the parameters of the original depression state prediction model (Text-ViT) are frozen, and only the MLP parameters are updated via training. By predicting the severity of depression in addition to the categorical state, the model provides richer information beyond simply classifying emotional text inputs. In the context of typical groups, a cautious approach is imperative when it comes to diagnosis and treatment, particularly in cases where individuals are on the borderline, as evidenced by symptoms that approximate mild depression. Consequently, when predicting the severity of depression, this tool is more robust in providing information than tools that merely indicate the presence of the condition.

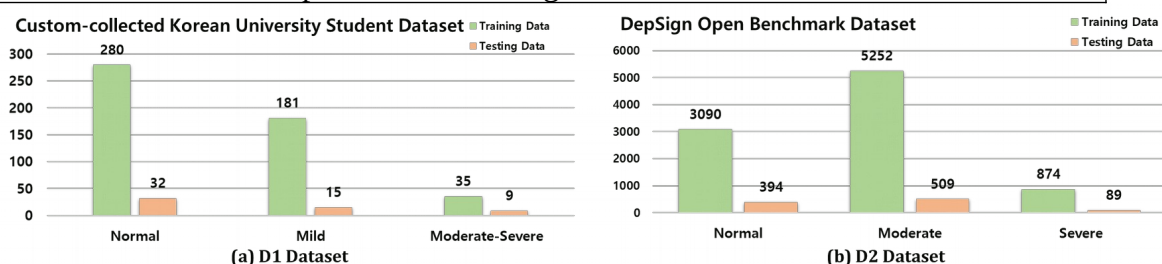
Results

Dataset and Experimental Setup

In this study, two datasets were utilized to conduct experiments on predicting the state and severity of depression based on text data. The first dataset (D1) is a Korean university student dataset explicitly collected for this study, and the second dataset (D2) is the publicly available DepSign [30] benchmark dataset.

The first dataset, D1, was collected from 30 undergraduate and graduate students currently enrolled at Sungkyunkwan University, which is the same dataset used in our previous work [41]. Out of a total of 630 data submissions, 44 missing entries and 34 short entries (less than 10 words) were excluded, resulting in a final dataset of 552 matched entries containing CES-D scores, PHQ-9 scores, and lifelog texts. Figure 7 (a) visualizes the distribution of “Normal, Mild, and Moderate-Severe” depression states among the 496 training and 56 testing samples used in the experiment. A majority of the data exhibits a bias toward the normal group. Conversely, the moderate to severe group comprises a small percentage of the total. This outcome can be attributed to the observation that the cohort of 30 subjects exhibited a distribution consistent with the general population and comprised individuals who were not currently afflicted with mental illness.

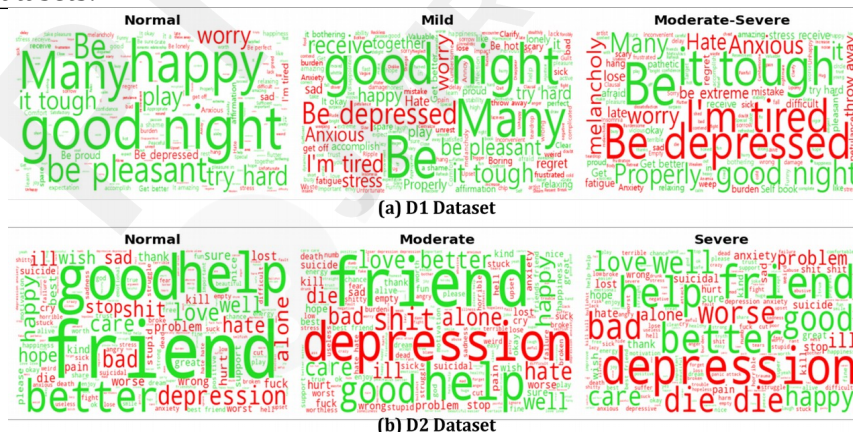
Figure 7. Distribution of depression state categories within the D1 and D2 dataset.



The second dataset (D2) was used to evaluate whether the proposed SD2Net can operate effectively across different cultural backgrounds. D2 is a depression-related dataset collected from Reddit, an online community platform similar in format to Twitter, using keyword-based crawling. Unlike D1, D2 features data drawn from a different cultural context, where English is the primary language of expression. In text-based lifelog data, where individuals describe their thoughts and emotions through language, the characteristics of the text can vary significantly depending on the language used and the writer's cultural background. East Asians tend to minimize and suppress their emotions more than Westerners, which is related to their cultural background of pursuing social harmony through individual emotional restraint [42]. Despite the diverse cultural backgrounds, we confirmed that the proposed SD2Net works robustly using D2. The GT labels in D2 consist of three categories: “Normal, Moderate, and Severe.” These were provided by trained experts based on the diagnostic criteria for Major Depressive Disorder (MDD) from the DSM-5. Figure 7 (b) visualizes the distribution of “Normal, Moderate, Severe” states in the training data and test data used in the experiment, showing that most of the data is biased toward the “Moderate” state.

We verified the differences in expression according to the cultural backgrounds of D1 and D2 through word clouds and polarity score analysis. Figure 8 shows the results of applying word clouds to the data for each depression state in the D1 and D2 datasets, where words with higher frequency appear in a larger font size. The color of each word is represented as “green” or “red,” depending on whether it is classified as “positive” or “negative” in the emotion lexicon. Notably, the D1 data was first analyzed in Korean and then translated into English only for visualization. In both datasets, the frequency of negative word usage increases as the state of depression increases, and particularly, expressions related to “depression” appear more frequently. Meanwhile, in D2, many words related to death are used and occupy a relatively large proportion, whereas in D1, references to death are rare. Instead, words such as “anxiety,” “worry,” and “fatigue” are frequently used across the entire group.

Figure 8. Word cloud analysis depicting the most frequent terms for each group within the D1 and D2 datasets.



Polarity score was calculated using the KNU Korean Sentiment Lexicon [43] for the D1 dataset and the VADER Lexicon [44] for the D2 dataset. A negative polarity value indicates negative sentiment, while a positive value indicates positive sentiment. For each data sample, the polarity values of all individual words were summed and then averaged to compute the sample-wise polarity score. Since the average of individual word scores is used, it is worth noting that this method does not account for contextual information. Therefore, polarity score analysis only allows for a superficial assessment of whether the words used in a sample are generally skewed toward negativity or positivity. The results

of polarity score analysis for the D1 dataset show that the means and standard deviations for the “Normal,” “Mild,” and “Moderate-Severe” states are 0.2422/0.5902, -0.1271/0.6051, and -0.4014/0.4913, respectively. In the D2 dataset, the corresponding values for the “Normal,” “Moderate,” and “Severe” states are -0.0811/0.7093, -0.3062/0.7376, and -0.3928/0.7407, respectively. A comparison of both datasets reveals disparities in the mean and variance between states of depression. However, in the case of D2, the mean polarity score for all states of depression is negative, indicating that most sentences are composed of negative words. This shows that, in general, the D2 dataset exhibits a greater prevalence of negative words compared to the D1 dataset.

All experiments were conducted on a workstation equipped with an Intel Core i9-13900K CPU, an NVIDIA RTX 4090 GPU with 24 GB of memory, and 128 GB of RAM. The software environment was based on Windows 10 with Python 3.8, PyTorch 1.13.1+cu116, and CUDA 12.6.

Results of HMHC Feature Index Extraction using GPT

Class values for each sub-item of the HMHC feature index were extracted from the raw dataset using a GPT model. This step is crucial for identifying HMHC-driven features that are medically meaningful, as it captures information from the overall contextual perspective of the raw text. While prior studies have typically extracted lexicon-based features to use text data as input for deep learning models, such approaches are limited in their ability to effectively capture medically relevant features due to the wide variability in expression and the lack of contextual understanding. In contrast, the HMHC feature extraction method proposed in the SD2Net enables the identification of features specifically tailored to mental health. In this study, we analyzed the class values of seven HMHC feature indices using the D1 and D2 datasets to evaluate the effectiveness of each feature index in distinguishing between different states of depression.

Figure 9 visualizes the distribution of class values for the MHI-9 feature index across different depression states in the D1 and D2 datasets. Figures 9 (a) to (c) present heatmaps for the “Normal,” “Mild,” and “Moderate-Severe” states in D1, respectively, while (d) to (f) correspond to the “Normal,” “Moderate,” and “Severe” states in D2. In the case of the D1 dataset, there are numerous instances in which the entire state selected the “Don't know” option. A comparison of Figures 9 (a) to (c) reveals a shift in the distribution of class values in a negative direction as the state of depression worsens. In the case of the D2 dataset, there is an overall tendency for class values to be classified as “Probable” or “Definite.” This pattern is commonly observed across all depression states—Normal, Moderate, and Severe—although, as shown in Figures 9 (d) to (f), there are slight variations in proportions among the states. In contrast, for the sub-items “Inability to Feel” and “Suicidal Thoughts,” it is evident that there is a distinct separation between the normal state and the other states.

Figure 9. Heatmap visualization of class values distributions for the MHI-9 feature index across different depression states in the D1 and D2 datasets: (a-c) D1 dataset, (d-f) D2 dataset.

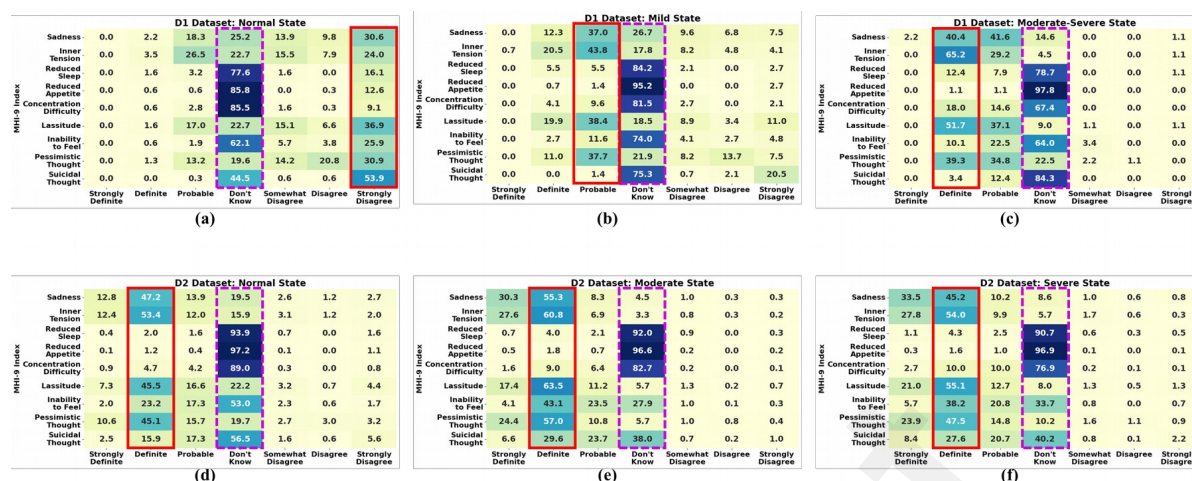


Figure 10 illustrates the distribution of class values for the ASI-16 feature index across different depression levels. Figures 10 (a) to (c) represent the D1 dataset, and Figures 10 (d) to (f) correspond to the D2 dataset. For the D1 dataset, Figures 10 (a) to (c) show that as the state of depression increases, the overall distribution shifts toward the “Negative” class values, while the range of the distribution becomes narrower. For the D2 dataset, its distributions are more heavily concentrated in the “Negative” range across all depression states. A comparison of (d) with (e) and (f) reveals a discrepancy in the distribution between the normal state and the subsequent states. Conversely, a comparison of (e) and (f) demonstrates an absence of statistically significant difference in distribution. Nevertheless, it is worth noting that for the same items, the proportion of “Heavily Negative” responses is slightly higher in the “Moderate” than in the “Severe”.

Figure 10. Heatmap visualization of class values distributions for the ASI-16 feature index across different depression states in the D1 and D2 datasets: (a-c) D1 dataset, (d-f) D2 dataset.

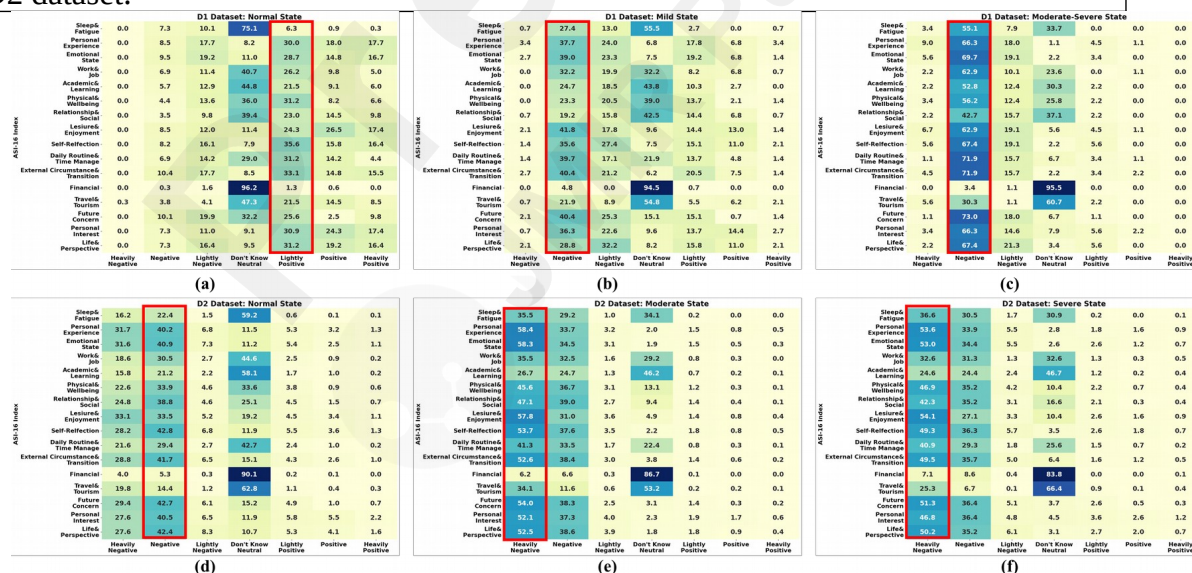


Figure 11 presents histograms showing the distribution of responses by depression state for the first sub-item of CDH-2 (“Experience of Depression Diagnosis”), as well as for AMH-1 and TWI-1. Figures 11 (a) to (c) correspond to the D1 dataset, while (d) to (f) correspond to the D2 dataset. Figures 11 (a) and (d) display the distribution of the “No” and “Yes” class values for the first sub-item of CDH-2 across depression states. Figures 11 (b) and (e) show the distributions for AMH-1, and Figures 11 (c) and (f) for TWI-1. As described, the D1 dataset consists of Korean university

students who have not experienced clinical depression or received a medical diagnosis of depression. Accordingly, for the medically related feature indices—CDH-2, AMH-1, and TWI-1—all data samples have a “No” class value. In contrast, the D2 dataset includes many individuals who have been diagnosed with depression or have taken antidepressant medication. Thus, in the distributions shown in Figures 11 (d) to (f), there is a shift in the proportions of “No” and “Yes” responses across depression states. For CDH-2, the proportion of “Yes” responses in the D2 dataset increases from 8.41% in the “Normal” state to 57.01% in the “Severe” state. Similarly, the proportion of “Yes” responses rises from 3.56% to 42.89% for AMH-1, and from 46.33% to 73.79% for TWI-1.

Figure 11. Class value distributions for the PDTH domain of HMHC (CDH-2, AMH-1, and TWI-1) across different depression states in the D1 and D2 datasets: (a-c) D1 dataset, (d-f) D2 dataset.

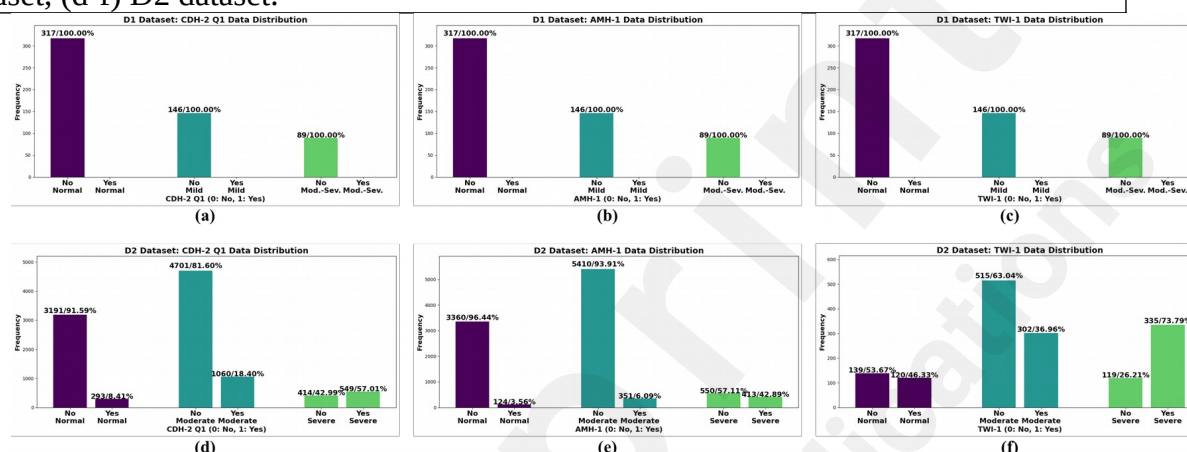


Figure 12 shows histograms of PH-1 scores by depression level. Figures 12 (a) to (c) correspond to the D1 dataset, while (d) to (f) correspond to the D2 dataset. PH-1 is defined as a score ranging from 1 to 10 that reflects the level of hope or perceived possibility of change regarding one's current situation. Scores closer to 1 indicate a more negative attitude, whereas scores closer to 10 indicate a more positive attitude. In Figures 12(a) to (c), the majority of data distributions for each state are characterized as follows: the Normal state scores six or higher, the Mild state ranges between three and seven, and the Moderate–Severe state ranges between two and four. In the D2 dataset, overall responses are concentrated in the 1-4 range. However, differences across depression states can be observed in the proportion of participants scoring under four and over seven: Normal (80.02%, 10.99%), Moderate (91.84%, 3.89%), and Severe (84.21%, 6.33%). In CDH-2, AMH-1, and TWI-1 results, a higher proportion of participants in the “Severe” state than in the “Moderate” state responded with “Yes,” which may reflect an increased willingness to cope with their situation through actions such as receiving a depression diagnosis or taking antidepressants.

Figure 12. Class value distributions for the PH-1 across different depression states in the D1 and D2 datasets: (a-c) D1 dataset, (d-f) D2 dataset.

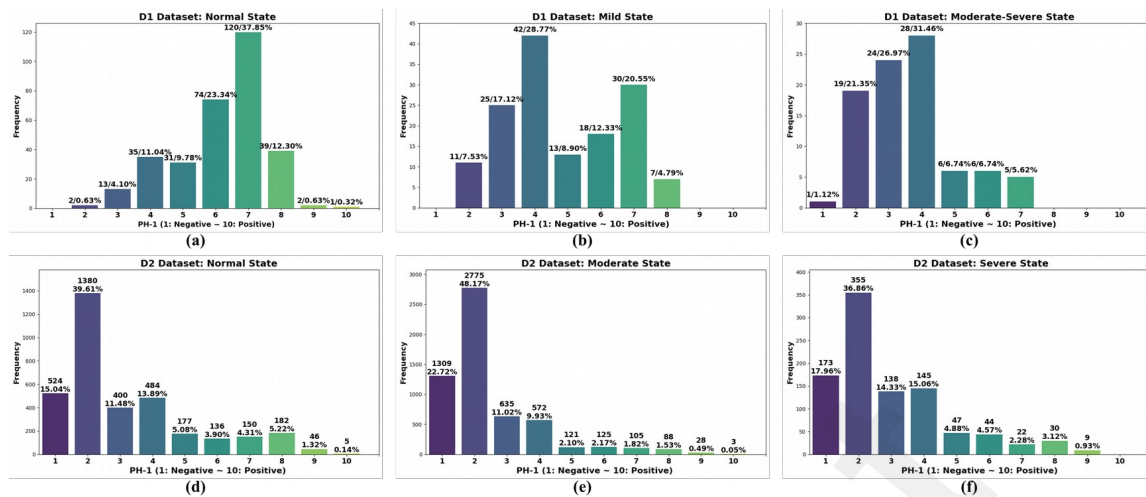
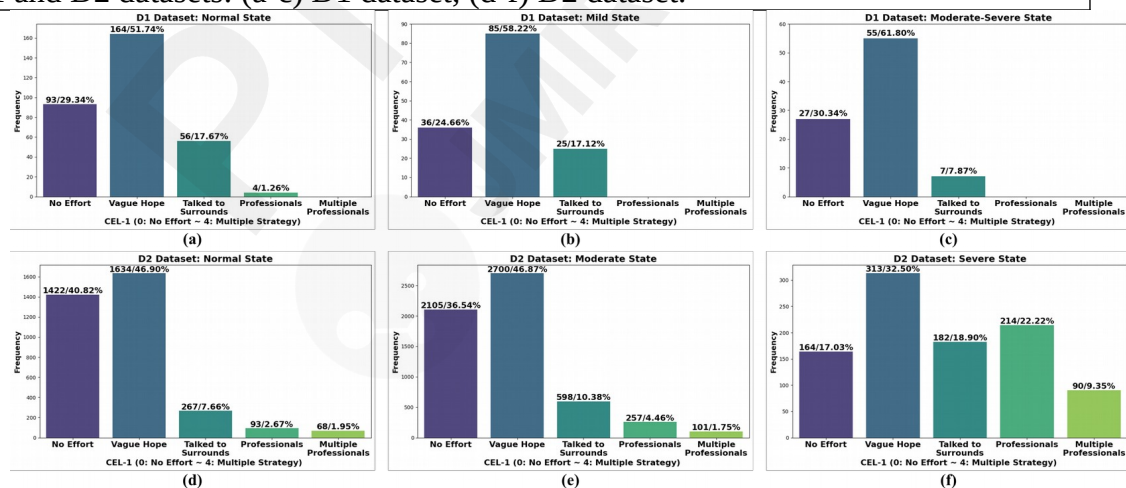


Figure 13. Class value distributions for the CEL-1 across different depression states in the D1 and D2 datasets: (a-c) D1 dataset, (d-f) D2 dataset.



Experimental Results of Depression State Prediction

Figure 14 presents the confusion matrices illustrating the prediction results of depression states on the D1 and D2 datasets using the proposed SD2Net. Despite the presence of class imbalance in both datasets, the prediction accuracy appears relatively well-balanced across the states. Table 3 presents the detailed evaluation results for depression state prediction on the D1 and D2 datasets, including the overall average accuracy, as well as state-wise accuracy, weighted precision, weighted recall, and

weighted F1-score. Weighted metrics were employed in this evaluation to account for class imbalance within the datasets and ensure that the performance assessment accurately reflects this imbalance.

Figure 14. Confusion matrices for depression state prediction: (a) D1 dataset, (b) D2 dataset.

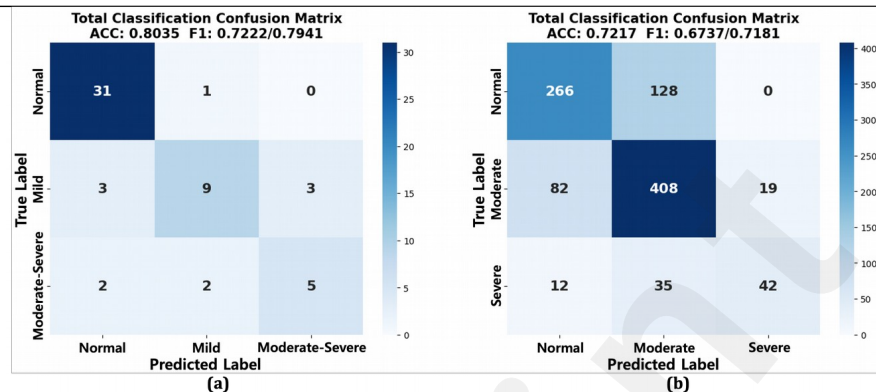


Table 3. The detailed performance of depression state prediction for the D1 and D2 datasets.

Dataset	Depression State	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
D1	Overall	80.36%	79.34%	80.36%	79.41%
	Normal	96.88%	86.11%	96.88%	97.18%
	Mild	60.00%	75.00%	60.00%	66.67%
	Moderate-Severe	55.56%	62.50%	55.56%	58.82%
D2	Overall	72.18%	72.19%	72.18%	71.82%
	Normal	67.51%	73.89%	67.51%	70.56%
	Moderate	80.16%	71.45%	80.16%	75.56%
	Severe	47.17%	68.85%	47.17%	56.00%

The D2 dataset is an open benchmark dataset that has been used in several prior studies [14, 16, 45, 46]. Table 4 compares the performance of the proposed SD2Net with the results reported in these previous studies. In terms of weighted F1-score, the proposed approach achieves the highest performance among all methods. It also demonstrates a notably high level of accuracy.

Table 4. Performance comparison with other studies on the D2 dataset.

Approach	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
SD2Net (Ours)	0.7218	0.7219	0.7218	0.7182
Ogunleye et al. [45]	0.74	0.71	0.68	0.69
Poświata et al. [16]	0.69	0.66	0.62	0.63
Muñoz et al. [14]	N/A	0.58	0.61	0.58
Thialakratne et al. [46]	0.663	0.656	0.623	0.635

Table 5 summarizes the comparative evaluation results of different baseline models on the D1 and D2 datasets, where the experiments were conducted using the same training and testing data. For BERT, RoBERTa, and DistilBERT baseline models, transfer learning was performed using the top four layers based on their pretrained weights. All experiments and tests were conducted on the same computing device that was used for SD2Net. In the case of GPT-4o, we applied a task-specific

prompt to classify depression severity, with the full prompt text provided in Multimedia Appendix 2.

For both the D1 and D2 datasets, the proposed SD2Net achieves higher performance in terms of accuracy, weighted precision, weighted recall, and weighted F1-score. In particular, concerning the weighted F1-score, the proposed approach outperforms the baselines by approximately 10% to 38% on the D1 dataset and by about 5% to 52% on the D2 dataset. These results indicate that the limitations inherent in conventional transformer-based approaches can be effectively overcome through the HMHC feature index, which is tailored to mental health, and the group-wise hierarchical model structure.

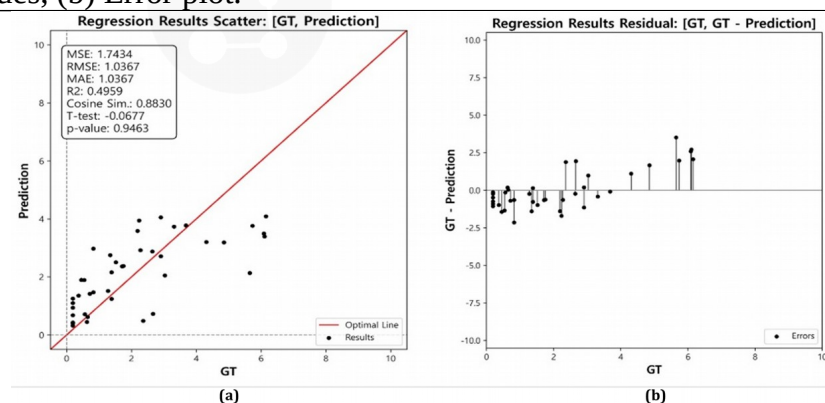
Table 5. Comparative evaluation of baseline transformer models on the D1 and D2 datasets.

Dataset	Baseline Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
D1	BERT	64.289%	66.59%	64.29%	59.67%
	RoBERTa	57.14%	32.65%	57.14%	41.56%
	DistilBERT	55.356%	76.54%	55.36%	56.07%
	GPT Prompt Engineering	69.09%	71.87%	69.09%	69.56%
	SD2Net (ours)	80.36%	79.34%	80.36%	79.41%
D2	BERT	67.04%	67.13%	67.04%	66.74%
	RoBERTa	66.84%	67.53%	66.84%	65.48%
	DistilBERT	65.93%	65.83%	65.93%	65.33%
	GPT Prompt Engineering	18.45%	52.06%	18.45%	19.67%
	SD2Net (ours)	72.18%	72.19%	72.18%	71.82%

Experimental Results for Depression Severity Prediction

Regarding depression severity, the two graphs in Figure 15 illustrate the GT of the test data and the regression results. The left graph illustrates the GT value on the x-axis and the prediction value on the y-axis. The right graph illustrates the error of the prediction value, with the x-axis representing the GT value and the y-axis denoting the (GT – prediction) value. In both graphs, the larger the GT value, the lower the prediction accuracy. This is interpreted to mean that the data used for training lacks sufficient samples with a “Moderate-Severe” level, resulting in a more significant error in the prediction.

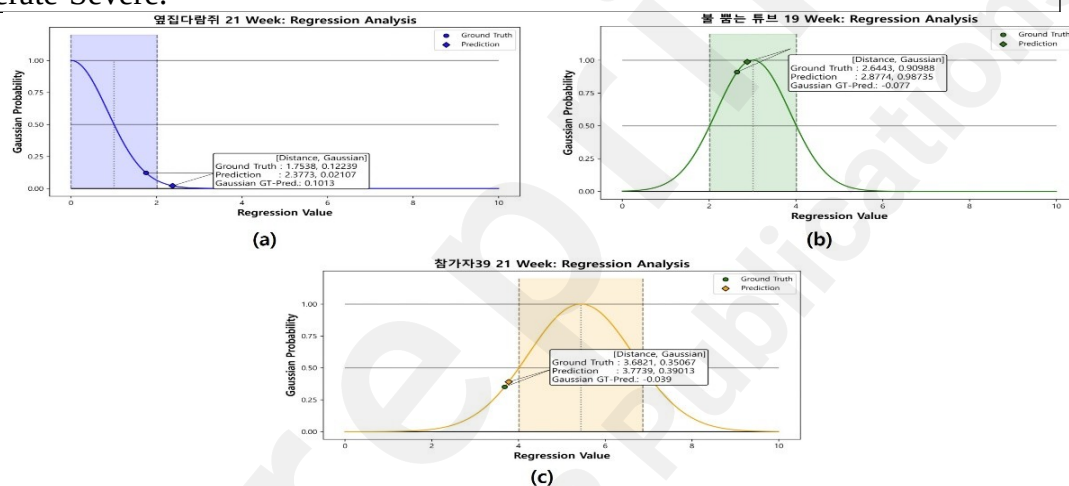
Figure 15. Performance of depression severity prediction: (a) Correlation between GT and predicted values, (b) Error plot.



We also examined the extent to which the depression severity prediction surpassed the depression

level prediction when utilized alone, employing a two-sample T-test analysis. The Gaussian probability of the GT depression level value and the predicted depression severity value in the Gaussian depression space was calculated. The absolute difference between the two values was determined to obtain the mean and standard deviation for the samples (mean: 0.3407, standard deviation: 0.2486). As a comparison experiment, we calculated the Gaussian probability of the depression level prediction result without the depression severity prediction result. The absolute value for the difference between the two values to obtain the mean and standard deviation for the samples (mean: 0.4929, standard deviation: 0.3577). Subsequently, a two-sample T-test was conducted, incorporating the means and standard deviations from those with depression severity and those without depression severity. The outcomes of this test indicated that the depression severity values exhibited a high degree of efficacy, with a 95% confidence level. Figure 16 presents a graphical representation of the test data's GT and prediction results. The circle points correspond to the GT, while the square points correspond to the prediction. When the GT lies on the boundary between levels, the corresponding depression severity can help clarify the ambiguity of the sample.

Figure 16. Examples of depression severity prediction results: (a) Normal, (b) Mild, (c) Moderate-Severe.



Ablation Study of Depression State Prediction

To verify the effectiveness of each component of the proposed SD2Net, we conducted an ablation study by removing or modifying its core modules. The ablation study targeted the following modules: Gaussian kernel encoding, HMHC with BERT (data-driven), group-wise clustering, and hierarchical structure. First, as shown in the first row of Table 6, replacing Gaussian kernel encoding with one-hot encoding resulted in approximately 4% (D1) and 5% (D2) lower weighted F1-scores compared with the proposed approach. Second, when only HMHC was used, the weighted F1-score decreased by 2% (D1) and 3% (D2). In contrast, when only BERT embeddings were used, the weighted F1-score dropped more substantially, by 26% (D1) and 4% (D2). Third, without group-wise clustering, the D1 dataset produced results identical to the proposed SD2Net because, as illustrated in Figures 12 (a) and (b), CDH-2 and AMH-1 values defined only a single group. In contrast, the D2 dataset allowed for clustering into four groups, and omitting group-wise clustering resulted in a decrease of approximately 7% in the weighted F1-score. Finally, without this hierarchical structure—i.e., directly predicting all three states in a single step—the weighted F1-score decreased by approximately 7%.

Table 6. Ablation study results for the proposed SD2Net components. All metrics are calculated using weighted values.

Target Module	Ablation	D1	D2
---------------	----------	----	----

	Configuration	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Gaussian Kernel Encoding	HMHC (One-hot) + Data-driven + Hierarchical + Group-wise	76.79%	76.38%	76.79%	75.17%	66.94%	67.36%	66.94%	66.51%
HMHC & BERT	HMHC (Gaussian) + Hierarchical + Group-wise	78.57%	76.93%	78.57%	77.34%	69.15%	70.21%	69.15%	68.49%
	Data-driven + Hierarchical + Group-wise	62.50 %	56.07%	62.50%	53.43%	68.25%	71.85%	68.25%	67.90%
Group-wise Clustering	HMHC (Gaussian) + Data-driven + Hierarchical	N/A	N/A	N/A	N/A	65.63%	66.24%	65.63%	64.71%
Hierarchical Structure	HMHC (Gaussian) + Data-driven + Group-wise	75.00%	75.20%	75.00%	70.83%	64.72%	64.96%	64.72%	64.23%
SD2Net (ours)	HMHC (Gaussian) + Data-driven + Hierarchical + Group-wise	80.36%	79.34%	80.36%	79.41%	72.18%	72.19%	72.18%	71.82%
Acc. = Accuracy; Prec. = Precision; Rec. = Recall.									

To evaluate the effectiveness of the newly introduced HMHC feature index for predicting depression states, we conducted an ablation study based on its four domains to examine which domains contribute most effectively. Experiments were performed using various combinations of feature indices from the four domains, and the results are summarized in Table 7. For the D1 dataset, the absence of HMHC feature indices from either the CDI or CTSF domain resulted in the most significant decrease in the weighted F1-score. In particular, when only a single domain among the four was used, the weighted F1-score dropped substantially. In contrast, the impact of the PDTH domain was minimal, which can be attributed to the characteristics of the D1 dataset, where there are no records of depression diagnosis or medication use, rendering this HMHC domain less meaningful. For the D2 dataset, all four domains contributed at a relatively similar level across different combinations; however, the absence of CDI and CTSF again resulted in the most significant decreases in the weighted F1-score. The PDTH domain was the next most influential, especially when two or more domains were moved simultaneously, in which case its contribution became more pronounced.

Table 7. Ablation study results with a combination of the four HMHC domains: CDI, CTSF, PDTH, and PPT. All metrics are calculated using weighted values.

HMHC Domain Configuration	D1				D2			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
CDI + CTSF + PDTH	78.57%	77.62%	78.57%	77.51%	63.27%	65.40%	63.27%	61.45%
CDI + CTSF + PPT	78.57%	78.10%	78.57%	77.81%	65.22%	67.90%	65.22%	63.68%
CDI + PDTH + PPT	75.00%	74.43%	75.00%	73.89%	62.50%	67.26%	62.50%	58.12%
CTSF + PDTH + PPT	76.79%	75.28%	76.79%	75.13%	63.72%	67.08%	63.72%	62.23%
CDI + CTSF	78.57%	78.44%	78.57%	78.32%	63.00%	66.34%	63.00%	60.46%
CDI + PDTH	75.00%	73.88%	75.00%	74.23%	62.45%	64.71%	62.45%	60.53%
CDI + PPT	78.57%	78.15%	78.57%	78.04%	63.51%	68.97%	63.51%	60.41%
CTSF + PDTH	78.57%	77.57%	78.57%	77.85%	64.21%	67.56%	64.21%	62.06%
CTSF + PPT	73.21%	72.65%	73.21%	71.02%	61.69%	65.62%	61.69%	58.46%
PDTH + PPT	75.00%	78.06%	75.00%	71.94%	64.92%	65.02%	64.92%	64.41%
CDI	75.00%	74.60%	75.00%	74.76%	69.29%	69.86%	69.29%	68.56%
CTSF	73.21%	72.36%	73.21%	72.07%	66.73%	66.77%	66.73%	66.28%

PDTH	57.14%	32.65%	57.14%	41.56%	66.43%	67.30%	66.43%	65.67%
PPT	71.43%	68.11%	71.43%	67.17%	58.97%	62.11%	58.97%	55.18%
SD2Net (ours)	80.36%	79.34%	80.36%	79.41%	72.18%	72.19%	72.16%	71.82%
<i>Acc. = Accuracy; Prec. = Precision; Rec. = Recall.</i>								

The ablation study on the components of SD2Net confirmed that each module plays a vital role in improving prediction accuracy. From the perspective of the D1 dataset, the HMHC feature index is shown to be particularly critical. In contrast, for the D2 dataset, the group-wise clustering and hierarchical structure learning strategies proved to be the most influential components. Furthermore, the effectiveness of the four HMHC domains tailored explicitly for mental health was validated. In particular, the CDI domain—defined based on clinical diagnostic criteria for mental health—and the CTSF domain—representing diverse topics and their corresponding sentiments—were found to play significant roles. In addition, depending on dataset characteristics, the PDTH domain, which reflects experiences of depression diagnosis and medication use, was also shown to contribute significantly to achieving high performance in depression state prediction.

Discussion

Principal Results

By extracting HMHC feature indices based on GPT as holistic clues and fusing the feature embedding of the HMHC feature indices with the lexicon-driven feature embedding through the proposed SD2Net, this study enables the prediction of multi-level depression states with their severity levels from naturally collected sentiment descriptions, thereby enhancing reliability and explainability. Furthermore, a learning strategy was implemented to construct a more effective and high-performance model. This strategy involved the preliminary grouping of samples based on high-dimensional information extracted through HMHC, followed by the construction of a hierarchical model prediction structure.

Limitations

In this paper, we employed both Korean and English datasets to substantiate the efficacy of SD2Net in transcending cultural boundaries. Although we achieved state-of-the-art performance in predicting multi-level depression states, further performance improvement is desirable for the depression self-prediction based on sentiment descriptions to play a key role in mental health self-monitoring platforms. Therefore, as future work, we plan to further enhance the performance of our proposed sentiment description-based depression prediction, e.g., by considering a history of data rather than data collected at a single point in time.

Comparison with Prior Work

The proposed SD2Net introduces a novel feature extraction, HMHC, which has not been attempted in prior studies. Additionally, it incorporates group-wise clustering and hierarchical model structure, which are training strategies for effective learning in a data-imbalanced environment. Especially, HMHC allows addressing the limitations of conventional lexicon-based or latent feature embedding-based approaches—such as their restricted capability of feature extraction from a clinical perspective and their focus on binary classification (non-depressed vs. depressed)—this study integrated a GPT-based HMHC feature index extraction module with group-wise clustering and a hierarchical structure, enabling more precise and higher-performing predictions. In particular, by leveraging the four domains of the HMHC feature index, the proposed approach was able to incorporate highly complex clinical information directly related to actual depressive symptoms, thereby contributing to predictions that are comparable to the level of clinical diagnosis. Also, the construction of an SD2Net was undertaken to predict not only the state of depression but also the severity of the current

depressive state. This approach yielded additional information on the depressive state for samples on the borderline. From the perspective of a mental health self-monitoring platform, SD2Net functions as an intermediary between individuals and psychiatrists, providing more robust information than approaches that merely predict the state of depression.

Conclusions

With the increasing prevalence of mental health disorders in the general population and the growing interest in self-monitoring platforms that enable non-face-to-face diagnosis and treatment, the proposed SD2Net can serve as a self-diagnosis module within such platforms. In particular, SD2Net allows users to predict their current depression state and severity level based on brief, 3–5 line sentiment descriptions, thereby eliminating the need for lengthy, multi-item self-report questionnaires. This allows individuals to self-assess their mental health state with high accuracy in a simple and accessible manner.

The proposed SD2Net introduces a novel feature extraction approach that has not been attempted in prior studies, leveraging GPT to extract the HMHC feature indices as a high-dimensional mental health indicator for holistic clues. This enables the effective handling of text data, which is inherently diverse and complex, from a mental health diagnosis perspective. In addition, to ensure that general contextual flow is also reflected in depression state prediction, SD2Net integrates lexicon-driven feature embeddings generated by BERT with HMHC feature embeddings generated by Text-ViT. Each feature index was probabilistically represented through Gaussian kernel encoding, thereby effectively capturing the inherent uncertainty in the data.

For predicting depression states, a hierarchical Text-ViT-BERT structure was employed to reduce problem complexity and maximize predictive accuracy. Furthermore, regression-based estimation of depression severity was incorporated, enabling not only state predictions but also the degree of severity within each stage, thereby supporting more fine-grained self-diagnosis. The severity scores, in particular, provide psychiatrists with richer and more informative references than categorical state alone, offering a more nuanced understanding of the user's condition.

In the future, we aim to advance the model beyond predicting current states by utilizing historical data to accomplish highly reliable mental health self-diagnoses. Finally, we plan to evaluate the developed self-diagnosis tools in terms of their clinical value, specifically regarding their role as an intermediary interface between the general population and psychiatrists within the framework of a self-monitoring platform.

Acknowledgements

SJL contributed to the experimentation, interpretation, writing, methodology, and data collection of the proposed approach. SHL contributed to the conceptualization, methodology, interpretation, and writing of the proposed approach.

This research was supported, in part, by the "Mental Health Monitoring: SNS-based Diagnosis and Analysis" Project (Grant No. HI22C0646) of the Korea Health Industry Development Institute (KHIDI), in part, by the "Disease Management Model Development Based on Smart Wearable Devices" Project (No. 2024ER080300) of the National Institute of Health (NIH), sponsored by the Korean Ministry of Health and Welfare (MOHW), and, in part, by Artificial Intelligence Graduate School Program (Sungkyunkwan University), Grant No. RS-2019-II190421, and by ICT Consilience Program, IITP-2020-0-01821, of the Institute of Information and Communication Technology Planning & Evaluation (IITP), sponsored by the Korean Ministry of Science and Information

Technology (MSIT).

Conflicts of Interest

None declared

Abbreviations

WHO: World Health Organization

UN: United Nations

MDD: Major Depressive Disorder

DSM-5: Diagnostic and Statistical Manual of Mental Disorders

LSTM: Long Short-Term Memory

RNN: Recurrent Neural Network

PCA: Principal Component Analysis

DistilBERT: Distilled Bidirectional Encoder Representations from Transformers

SqueezeBERT: Squeezed-and-Excitation Bidirectional Encoder Representations from Transformers

DeBERTa: Decoding-enhanced Bidirectional Encoder Representations from Transformers with Disentangled Attention

RoBERTa: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach

GPT-3.5: Generative Pre-trained Transformer 3.5

GPT-4: Generative Pre-trained Transformer 4

LLM: Large Language Model

PHQ-9: Patient Health Questionnaire-9

BSSI: Beck Scale for Suicide Ideation

MAE: Mean Absolute Error

PHQ-8: Patient Health Questionnaire-8

ML: Machine Learning

DsD: Depression Sign Dataset

DepSign: DepSign Depression Dataset

BERT: Bidirectional Encoder Representations from Transformers

GT: Ground Truth

MLP: Multi-layer Perceptron

CES-D: Center for Epidemiologic Studies Depression Scale

VAD: Valence, Arousal, Dominance

ViT: Vision Transformer

VADER: Valence Aware Dictionary and sEntiment Reasoner (VADER) Lexicon

References

1. United Nations. (2022). *COVID-19's impact on mental health*. UN Chronicle. Retrieved October 4, 2025, from <https://www.un.org/en/chronicle/covid-19s-impact-mental-health>
2. United Nations. (2022). *COVID-19's impact on mental health*. UN Chronicle. Retrieved October 4, 2025, from <https://www.un.org/en/chronicle/covid-19s-impact-mental-health>
3. Dozois, D. J. (2021). Anxiety and depression in Canada during the COVID-19 pandemic: A national survey. *Canadian Psychology/Psychologie canadienne*, 62(1), 136.
4. Kim, K. Y., Yang, Y. B., Kim, M. R., Kim, J., & Park, J. S. (2025). Predicting early depression in WZT drawing image based on deep learning. *Expert Systems*, 42(2), e13675.
5. Tønning, M. L., Faurholt-Jepsen, M., Frost, M., Martiny, K., Tuxen, N., Rosenberg, N., ... & Kessing, L. V. (2021). The effect of smartphone-based monitoring and treatment on the rate

- and duration of psychiatric readmission in patients with unipolar depressive disorder: the RADMISS randomized controlled trial. *Journal of Affective Disorders*, 282, 354-363.
6. Kinoshita, S., Cortright, K., Crawford, A., Mizuno, Y., Yoshida, K., Hilty, D., ... & Kishimoto, T. (2022). Changes in telepsychiatry regulations during the COVID-19 pandemic: 17 countries and regions' approaches to an evolving healthcare landscape. *Psychological medicine*, 52(13), 2606-2613.
 7. Dogan, E., Sander, C., Wagner, X., Hegerl, U., & Kohls, E. (2017). Smartphone-based monitoring of objective and subjective data in affective disorders: where are we and where are we going? Systematic review. *Journal of medical Internet research*, 19(7), e262.
 8. Lee, S., Lee, S., & Lee, Y. (2024, August). Self-Monitoring the Mental-Health State of a Focused Population with Multiple Self-Questionnaires and Sentiment Descriptions. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 674-680). IEEE.
 9. Zafar, A., Aftab, D., Qureshi, R., Wang, Y., & Yan, H. (2024). Multi-Explainable TemporalNet: An Interpretable Multimodal Approach using Temporal Convolutional Network for User-level Depression Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2258-2265).
 10. Amanat, A., Rizwan, M., Javed, A. R., Abdelhaq, M., Alsaqour, R., Pandya, S., & Uddin, M. (2022). Deep learning for depression detection from textual data. *Electronics*, 11(5), 676.
 11. Bokolo, B. G., & Liu, Q. (2023). Deep learning-based depression detection from social media: Comparative evaluation of ml and transformer techniques. *Electronics*, 12(21), 4396.
 12. Shin, D., Kim, H., Lee, S., Cho, Y., & Jung, W. (2024). Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: instrument validation study. *Journal of Medical Internet Research*, 26, e54617.
 13. Milintsevich, K., Sirts, K., & Dias, G. (2023). Towards automatic text-based estimation of depression through symptom prediction. *Brain informatics*, 10(1), 4.
 14. Muñoz, S., & Iglesias, C. Á. (2023). Detection of the severity level of depression signs in text combining a feature-based framework with distributional representations. *Applied Sciences*, 13(21), 11695.
 15. Kabir, M., Ahmed, T., Hasan, M. B., Laskar, M. T. R., Joarder, T. K., Mahmud, H., & Hasan, K. (2023). DEPTWEET: A typology for social media texts to detect depression severities. *Computers in Human Behavior*, 139, 107503.
 16. Poświata, R., & Perełkiewicz, M. (2022, May). Opi@ It-edi-acl2022: Detecting signs of depression from social media text using roberta pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 276-282).
 17. Kim, S., Imieye, O., & Yin, Y. (2025). Interpretable Depression Detection from Social Media Text Using LLM-Derived Embeddings. *arXiv preprint arXiv:2506.06616*.
 18. American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders*. American psychiatric association.
 19. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
 20. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
 21. Iandola, F. N., Shaw, A. E., Krishna, R., & Keutzer, K. W. (2020). SqueezeBERT: What can computer vision teach NLP about efficient neural networks?. *arXiv preprint arXiv:2006.11316*.
 22. He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
 23. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A

- robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
24. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12), 2009.
 25. Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9), 606-613.
 26. Beck, A. T., Kovacs, M., & Weissman, A. (1979). Assessment of suicidal intention: the Scale for Suicide Ideation. *Journal of consulting and clinical psychology*, 47(2), 343.
 27. Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3), 163-173.
 28. Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., ... & Morency, L. P. (2014, May). The distress analysis interview corpus of human and computer interviews. In *LREC* (Vol. 14, pp. 3123-3128).
 29. Naseem, U., Dunn, A. G., Kim, J., & Khushi, M. (2022, April). Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM web conference 2022* (pp. 2563-2572).
 30. Sampath, K., & Durairaj, T. (2022, March). Data set creation and empirical analysis for detecting signs of depression from social media postings. In *International Conference on Computational Intelligence in Data Science* (pp. 136-151). Cham: Springer International Publishing.
 31. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).
 32. Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10), e22635.
 33. Boinepelli, S., Raha, T., Abburi, H., Parikh, P., Chhaya, N., & Varma, V. (2022, June). Leveraging mental health forums for user-level depression detection on social media. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 5418-5427).
 34. Garg, M., Saxena, C., Krishnan, V., Joshi, R., Saha, S., Mago, V., & Dorr, B. J. (2022). Cams: An annotated corpus for causal analysis of mental health issues in social media posts. *arXiv preprint arXiv:2207.04674*.
 35. Priyadarshana, Y. H. P. P., Liang, Z., & Piumarta, I. (2023, August). Heladepdet: A novel multi-class classification model for detecting the severity of human depression. In *International Conference on Collaboration Technologies and Social Computing* (pp. 3-18). Cham: Springer Nature Switzerland.
 36. Rani, S., Ahmed, K., & Subramani, S. (2024). From posts to knowledge: annotating a pandemic-Era Reddit dataset to navigate mental health narratives. *Applied Sciences*, 14(4), 1547.
 37. Rahman, A. B. S., Ta, H. T., Najjar, L., Azadmanesh, A., & Gönul, A. S. (2024). DepressionEmo: A novel dataset for multilabel classification of depression emotions. *Journal of Affective Disorders*, 366, 445-458.
 38. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
 39. OpenAI. GPT-3.5 Technical Report. OpenAI, 2023, <https://openai.com/research/gpt-3-5>.
 40. Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the

- general population. *Applied psychological measurement*, 1(3), 385-401.
41. Lee, S., Lee, S., Lee, J., Jo, Y. T., Park, E., & Cha, J. (2024). Fusion of multiple self-diagnostic questionnaires into optimal diagnostic cut-offs and factor analysis for depression characterization of the Korean university student group. *BMC psychiatry*, 24(1), 820.
 42. Song, H., Chan, J. S., & Ryan, C. (2024). Differences and similarities in the use of nine emotion regulation strategies in Western and East-Asian cultures: Systematic review and meta-analysis. *Journal of Cross-Cultural Psychology*, 55(8), 865-885.
 43. Park, S. M., Na, C. W., Choi, M. S., Lee, D. H., & On, B. W. (2018). KNU Korean sentiment lexicon: Bi-LSTM-based method for building a Korean sentiment lexicon. *Journal of Intelligence and Information Systems*, 24(4), 219-240.
 44. Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).
 45. Ogunleye, B., Sharma, H., & Shobayo, O. (2024). Sentiment Informed Sentence BERT-Ensemble Algorithm for Depression Detection. *Big Data and Cognitive Computing*, 8(9), 112.
 46. Thialakratne, S., Welgama, V., & Weerasinghe, R. (2024). Detecting the Severity of Depression in Online Forum Data by Leveraging Implicit Semantic Inferences. *International Journal on Advances in ICT for Emerging Regions (ICTer)*, 17(2).

Supplementary Files

Multimedia Appendixes

HMHC feature indices configurations.

URL: <http://asset.jmir.pub/assets/03892d68013ee31774b2777ec096456d.docx>

GPT-4o prompt for depression state prediction (comparative evaluation study).

URL: <http://asset.jmir.pub/assets/30b812150ee265679025ef995220be65.docx>