Chula Med J Vol. 68 No. 4 October - December 2024 ; 261 - 273 DOI : 10.56808/2673-060X.5495

Original article

Accuracy and concordance of DMIND AI application with a renowned depression assessment tool in Thai adults

Solaphat Hemrungroj[a, b, c, d, *], Kittipoch Saengsai[a], Pasit Jakkrawankul[a], Chanyanart Kiattiporn-Opas[a], Kantapat Chaichareenon[a], Arisara Amrapala[a, b, d], Kulvara Lapanan[e, f], Titipat Achakulvisut[f], Peerapol Vateekul[g], Natawut Nupairoj[a, h], Phanupong Phutrakool[c, i, j], Rapinpat Yodlorchai[a], Narin Hiransuthikul[c], Sarunya Hengprapom[c]

[a]*Center of Excellence in Digital and AI for Mental Health, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand*
[b]*Cognitive Fitness and Bio Psychiatry Technology Research Unit, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand*
[c]*Department of Preventive and Social Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand*
[d]*Department of Psychiatry, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand*
[e]*Neuroscience Research Australia, Randwick, New South Wales, Australia*
[f]*School of Population Health, Faculty of Medicine and Health, University of New South Wales, Sydney, New South Wales, Australia*
[g]*Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, Bangkok, Thailand*
[h]*Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand*
[i]*Chula Data Management Center, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand*
[j]*Center of Excellence in Preventive and Integrative Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand*

Abstract

Background: In Thailand, the high prevalence of depression, particularly in rural areas with limited mental health services, poses significant challenges. Advanced technologies like digital phenotyping and artificial intelligence (AI), particularly natural language processing (NLP) and Large Language Models (LLMs), offer promising solutions by analyzing digital data to detect early signs of depression.

Objectives: This study evaluated the performance of the "Detection and Monitoring Intelligence Network of Depression (DMIND)" application, an AI-powered screening tool for detecting depression in Thai adults that analyzes behavioral data from participant responses using machine learning algorithms, such as NLP models and LLMs. The study aimed to determine the agreement of the DMIND AI model and the Thai version of the Hamilton depression rating scale (HDRS-17 Thai) in classifying depression severity.

Methods: This cross-sectional study recruited 388 participants from one tertiary care hospital and two psychiatric hospitals. Initially, participants used the DMIND application, where they were asked to answer a series of questions. Their response in the application was recorded and a pre-trained AI model predicted their depression severity. Subsequently, a trained nurse or psychologist then assessed participants using the HDRS-17 Thai to establish a baseline measure of depression severity. Statistical analysis involved comparing the depression severity classifications from the DMIND AI model with the HDRS-17 Thai. Cohen's kappa coefficient, sensitivity, specificity, and predictive values were used to evaluate the agreement between the two assessments.

Results: Our DMIND application demonstrated moderate agreement with the HDRS-17 Thai, indicating substantial consistency in depression severity classification. The tool showed high sensitivity (87.3%) and moderate specificity (59.5%), with strong negative predictive values for detecting depression.

Conclusion: The AI-powered DMIND application effectively screens for depression by analyzing digital data from participants' responses. Its moderate agreement with a traditional clinical assessment and strong diagnostic performance highlights its potential as a scalable, accessible tool for mental health management in Thailand. Integrating AI tools like the DMIND into the public health infrastructure could significantly enhance the accessibility, accuracy, and responsiveness of mental health services, particularly in underserved regions, potentially revolutionizing the management and treatment of depression across the country.

Keywords: Artificial intelligence, depression, detection.

***Correspondence to: Solaphat Hemrungrojn,** Center of Excellence in Digital and AI for Mental Health, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand.
E-mail: Solaphat.h@chula.ac.th
Received: June 15, 2024
Revised: June 4, 2024
Accepted: July 17, 2024

Major depressive episodes (MDE) encompass a spectrum of symptoms, from insomnia to hypersomnia, and from weight loss to increased appetite, highlighting the complexity and variability of depression. Globally, nearly 5.0% of the population experiences depression annually, with severe cases often linked to increased suicide risks.[1] Recent global events, including the COVID-19 pandemic, have intensified the need for effective depression management strategies encompassing prevention, timely diagnosis, and comprehensive treatment.[2]

In Thailand, the prevalence of depression poses significant challenges, especially in rural areas where access to mental health services is limited. With only 719 psychiatrists serving nearly 3 million individuals with depressive conditions, the country faces significant hurdles in mental health diagnosis and treatment.[3] This disparity underscores an urgent need for enhanced mental health infrastructure and services to sufficiently support affected populations.

The traditional approach to diagnosing depression in Thailand involves extensive one-on-one consultations, which are impractical on a large scale due to the scarcity of trained mental health professionals. During crises, such as severe depressive episodes or suicidal ideations, the demand for psychiatric help surges, yet the availability of professionals cannot keep pace, often leading to inadequate care. Traditional diagnostic methods, such as the Diagnostic and statistical manual of mental disorders, fifth edition, text revision (DSM-5 TR) and the 17-item Hamilton depression rating scale (HDRS-17), while widely used, are time-consuming and limited by the availability of trained professionals. On the other hand, self-reported assessments like the patient health questionnaire-9 (PHQ-9) may be influenced by the patient's state of mind or ability to accurately report symptoms. Importantly, these traditional approaches may struggle to capture the nuanced and dynamic nature of depressive symptoms.

As an attempt to combat problems with access and time, the Thai government set up a 1323 mental health hotline for the public. Nevertheless, the hotline has seen an overwhelming increase in calls beyond the capacity of available staff, emphasizing the current system's inability to meet this demand effectively and the urgent need for efficient mental health services.

In this challenging context, technology, particularly digital phenotyping and artificial intelligence (AI), offers a beacon of hope. By leveraging data from smartphones and other personal devices, digital phenotyping can monitor and diagnose depression by analyzing changes in physical activity, social interactions, and overall device usage. These technologies help bridge the gap in mental health services and offer scalable solutions particularly beneficial in remote or underserved regions. This potential method of passive data collection can help detect early signs of depression with minimal intrusion, representing a significant advancement in mental health management.

The role of advanced AI technologies such as Natural language processing (NLP) and Large language models (LLMs) becomes indispensable, especially in time-sensitive scenarios. These technologies can significantly enhance the efficiency and reach of mental health services. By analyzing language patterns and detecting nuances in communication, NLP and LLMs can help identify the severity of depression in real time. One example of the utilization of AI models is a study by Moreno MA, *et al.*[4] where personally written text from public Facebook profiles of college students were evaluated with the DSM-5 criteria for depression symptoms or MDE. Coders were trained using the DSM-5 criteria and negative binomial regression was used to model the associations. Results indicated that 25.0% of profiles displayed depressive symptoms, with 2.5% meeting the MDE criteria. Common symptoms identified included depressed mood, guilt or worthlessness, indecisiveness, and loss of energy. Factors such as recent Facebook activity and receiving responses to depressive disclosures were associated with increased numbers of symptom presentations. Similarly, Choudhury MD, *et al.*[5-7] identified a set of Twitter users diagnosed with clinical depression, measured their social media behavior, and built a statistical classifier to estimate depression risk. The study found that certain social media signals, such as decreased social activity, could characterize the onset of depression. The classifier predicted depression with 70.0% accuracy.

Apart from text, speech, which integrates neuromuscular, physiological, and cognitive elements, has emerged as a potential biomarker for mental disorders like mild depressive disorder (MDD) and can be analyzed with computational techniques. For example, the RADAR-MDD multi-center explored the use of speech as a biomarker for MDD by

collecting speech data from 585 participants with a history of MDD in the UK, Spain, and the Netherlands. Over 18 months, participants recorded their speech bi-weekly on their smartphones. The authors then analyzed 28 speech features and used linear mixed models to determine their association with depressive symptoms. Their results found that increased depressive symptoms were significantly correlated with measures such as speech rate and articulation rate.

Considering, the extensive capability of NLP models and LLMs, these methods would not only aid in triaging cases and prioritizing those in immediate need but will also support mental health professionals by providing detailed, data-driven insights into patient conditions without needing direct human intervention in every case. Integrating these technologies into services like the 1323 hotline could dramatically improve response times and the quality of care, potentially saving lives by bridging the gap between the demand for mental health services and the availability of professional help.

Previously, our study group developed and validated the "Detection and monitoring intelligence network of depression (DMIND)" questionnaire, an alternative depression screening tool for the Thai population that was accurate and capable of AI integration. [7] We incorporated this questionnaire into an application (DMIND application) where the assessment was administered via an avatar resembling a psychiatrist. Afterward, we conducted several pilot tests and developed and refined an AI scoring model for depression, which has been integrated into the application. The main objective of this study was to evaluate the performance of our AI model in the Thai population. We hope to transform Thailand's mental health care landscape by encouraging the development and utilization of AI tools such as the DMIND application.

Materials and methods

Study design

This is a cross-sectional study evaluating the performance of the AI-powered DMIND application in depression detection and classification. The AI model uses a combination of NLP and LLMs to analyse responses from a previously developed clinical assessment adapted for the Thai population. This study was conducted ethically following the World Medical Association Declaration of Helsinki. The study was approved by the Ethical Review Board of Somdet Chaopraya Institute of Psychiatry (008/2566), the Faculty of Medicine, Chulalongkorn University (COA no. 1266/2023), and Srithanya Hospital (017/2566).

Participant selection

Participants were recruited from outpatient departments at Somdet Chaopraya Institute of Psychiatry, King Chulalongkorn Memorial Hospital and Srithanya Hospital in Thailand. To ensure a representative mix, we used purposive sampling to include a diverse range of normal controls and patients diagnosed with depression. Eligible participants were adults aged 18 to 65 years who were fluent in Thai, had no intellectual disability, and consented to video and audio recording. Individuals with severe cognitive impairment, psychotic disorders, inability to communicate, or facial expression issues due to conditions like Bell's palsy or Parkinson's disease were excluded.

Sample size calculation

The sample size was calculated based on the work of Obuchowski NA. [8], which provides a formula for calculating sample size in studies examining test accuracy. With an expected agreement rate of 80.0%, a confidence level of 95%, and a power of 80.0%, the required sample size was 354 participants. This size ensures that the study is adequately powered to detect significant agreement between the DMIND application and traditional clinical assessments.

Data collection

After providing consent, participants used the DMIND application where they were asked to answer a series of questions. Participants provided answers through video conversation with a psychiatrist avatar. Their response was recorded in the application and the AI model predicted their depression severity. Subsequently, after some delay, trained clinicians or nurses then assessed participants using the HDRS-17 to establish a baseline measure of depression severity. The delayed administration was designed to prevent any bias or influence in responses between the two assessments and ensure the integrity of the comparative analysis.

Screening tools

DMIND application

The complete DMIND application

The DMIND application is an AI-assisted application that automatically assesses the user's depressive score. This application, along with its components, has been continuously developed and improved since 2020, and was first released in 2022 under the Center of Excellence in Digital and AI for Mental Health at Chulalongkorn University. **Figure 1** illustrates how patients use the DMIND application. The application can be accessed from various channels, such as the Mohprompt Line official account. Once inside, patients can be asked for consent and permission to turn on their cameras and microphone before being interviewed by a psychiatrist avatar. After completing the assessment, the patient's depression risk level (green: normal, yellow: mild, orange: moderate and red: severe) is displayed, whereby this risk level is determined by AI analysis of the interview responses. Those with a severe or moderate depression risk level can voluntarily provide their mobile number in the application for a psychiatrist to contact them afterward to provide them with the support they need.

Prior development of the DMIND AI depression scoring model

Previously, the authors developed two depression scoring supervised learning models: 1) a scoring model from call records from the Department of Mental Health's (DMH's) 1323 Mental Health Hotline; and 2) a scoring model from interview videos of patients at King Chulalongkorn Memorial Hospital, Thai Redcross Society. However, both scoring models were built upon the full version of well-established depression screening tools (PHQ-9 and HDRS-17 respectively), which are time-consuming as they involve many questions. Therefore, further improvement of the questionnaire and scoring model was needed.

Following the development of the DMIND questionnaire and its incorporation into our DMIND application in a previous study [7] new testing data was collected in a subsequent study and the previous AI models were combined and adjusted for the context of the new questionnaire. At least 2 - 4 psychiatrists evaluated and provided the severity level labels for each interview session in the test dataset. For this updated model, various modern deep learning models were applied: 1) DeepFace [9] was used for extracting facial expression; 2) ThaiSER was used for extracting voice sentiment; 3) TwHIN-BERT [10] was chosen as our pre-trained model for textual data; and 4) Thonburian whisper was our automatic speech recognition (ASR) model for transcribed text. From these deep learning models, the combined depression scoring model was developed and many different combinations of multimodal features **(Figure 2)** were selected and evaluated. A multimodal model for each item in the DMIND questionnaire was constructed and then resembled using regression.

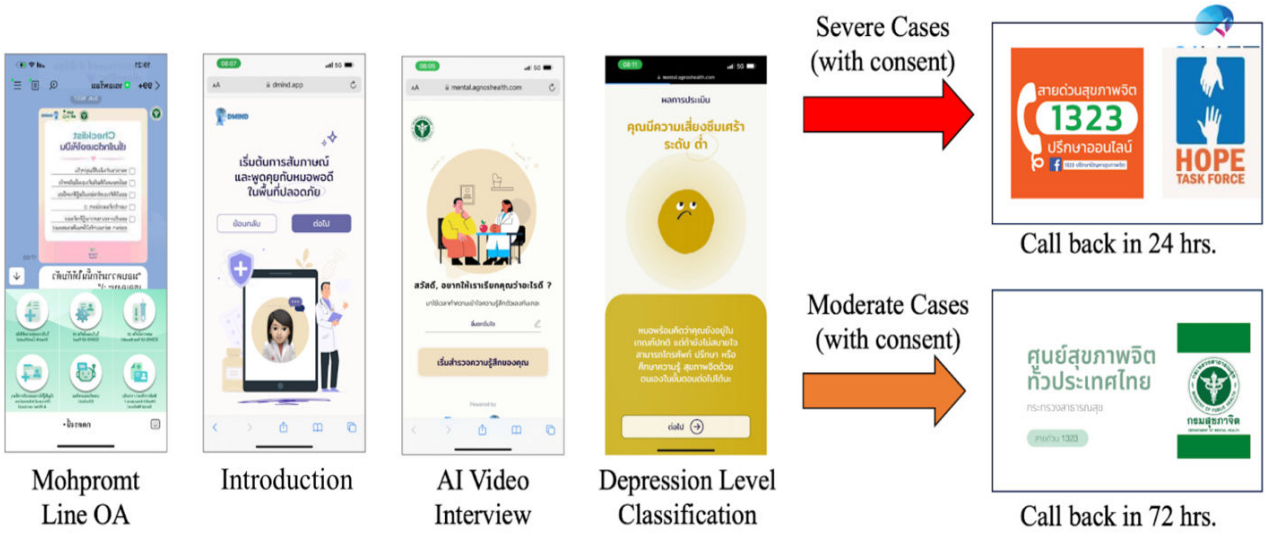

Figure 1. How patients use the DMIND application.

The multimodal model performance was evaluated using testing data collected from July 2022 to December 2022. Various combinations of modalities from interview videos were examined, including textual only, textual+audio, and textual+audio+video, to classify individuals into depressive and non-depressive categories. The F1 score, which balances precision and recall, showed no significant difference among the three combinations (textual only: 0.7777, textual+audio: 0.7752, and textual+audio+facial: 0.7726).

Upon further trials and development, we encountered significant challenges. A notable percentage of participants were unwilling to give permission to open camera during interview. Additionally, controlling participants to consistently position their faces within the frame proved to be difficult, with most of them failing to comply. Considering these constraints alongside the results from the process development phase, we concluded that utilizing the textual-only approach, which demonstrated the highest F1 score of 0.7777, was the most viable option. This textual-only model was incorporated into the final DMIND application.

Dataset, input, and target

As of November 14, 2022, there are a total of 1,252 labelled records that make up the 'DMIND dataset', which was used to continuously train all the AI models up until the final model used in this study.

The DMIND application scores the depression severity level from interview response videos. Different AI technologies are incorporated into this application, such as deep learning, generative AI, ASR, NLP, and LLMs. To achieve the highest performance, the final DMIND AI model was comprised of two main models, as shown in **Figure 3**: 1) a binary classification for depressive scores based on a

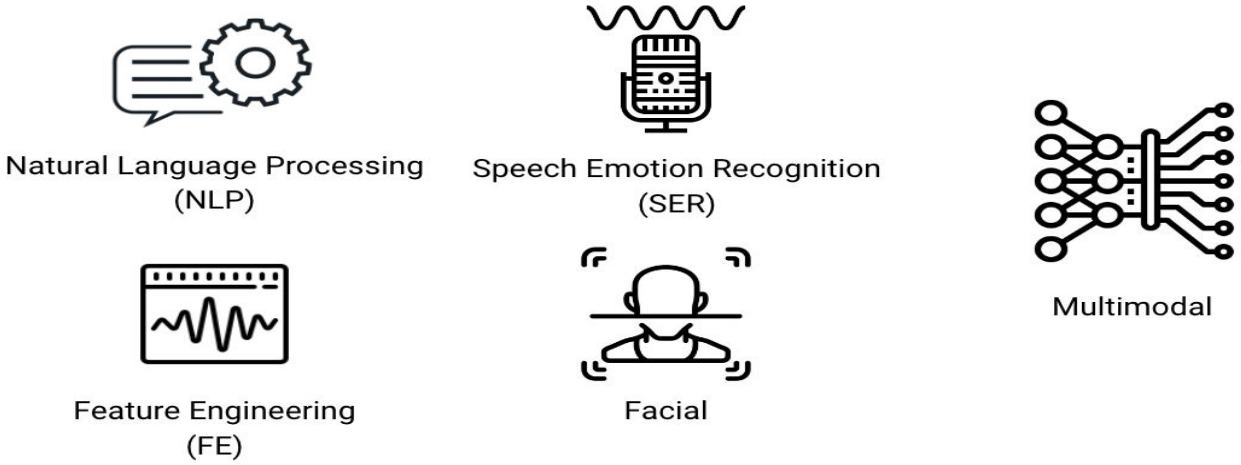

Figure 2. The multimodal model utilizes all available information from the interview videos, e.g., textual content (NLP), voice sentiment (SER), voice feature (FE), and facial expression.

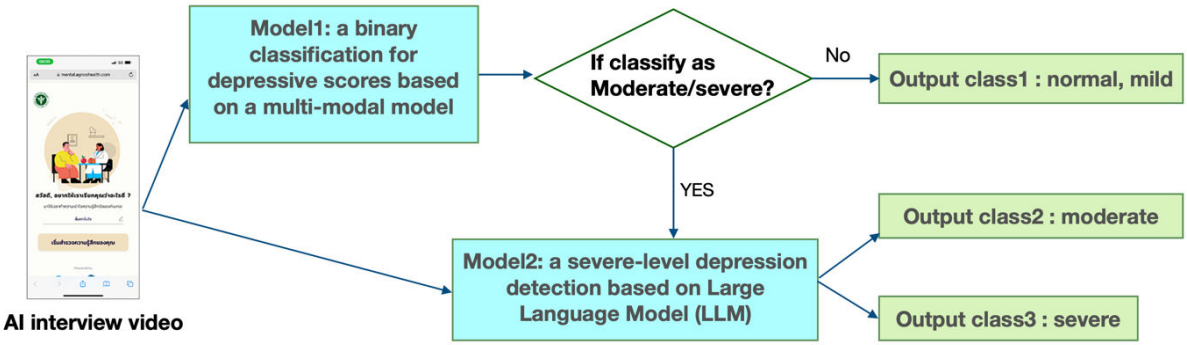

Figure 3. The final DMIND AI model.

multimodal model and 2) a severe-level depression detection based on LLM. Accordingly, the DMIND application classifies the depressive score of each patient into three classes: 1) normal/mild; 2) moderate; and 3) severe.

Model 1: Binary classifier development

As the first step in the system, the multimodal model aims to classify depression into binary classes: normal/mild and moderate/severe. There are two main modalities: voice and textual content. Facial features and expressions were not included since many patients did not turn on their cameras. For voice data, Thonburian Whisper was chosen as the ASR model. The Thonburian Whisper is derived from Open AI's Whisper and was refined using numerous Thai datasets. Additionally, we also fine-tuned the Thonburian Whisper with over 10 hours of our speech data transcribed from the DMIND application, enhancing its ability to accurately transcribe text in the mental health context.

To further analyze the textual content from the speech transcription, TwHIN-BERT was selected as our pre-trained model for textual data. The TwHIN-BERT is an innovative multi-lingual tweet language model trained on 7 billion tweets from more than 100 languages, including Thai. This model also incorporates a social objective that leverages the extensive social interactions within a Twitter Heterogeneous Information Network (TwHIN). TwHIN-BERT was further improved and fine-tuned on our training data to classify depression into two output levels: normal/mild and moderate/severe.

Model 2: LLM Development

Only moderate and severe patients from the first model proceed to the second model (as shown in **Figure 3**), a severe-level depression detection based on LLM. Here, we employed a commercial LLM, e.g., Chat GPT. A key sign of a severe depressive level is a suicidal attempt. Thus, we designed a prompt in the LLM to classify transcribed text into seven suicidal levels, namely level 0 : no suicidal ideation at present; level 1 : occasional thoughts of death; level 2 : passive death wish; level 3 : suicidal ideation; level 4 : attempted suicide in the past; level 5 : persistent suicidal ideation with high frequency; level 6 : self-injury; and level 7 : suicidal plan or attempt at present. Patients with suicidal levels less than 4 were considered non-severe.

Final model performance

To evaluate performance, the testing data was separated from the training data. For the first model (normal/mild vs. moderate/severe), the results showed that it achieved an area under the receiver operating characteristic (ROC) curve of 87.0%, an F1 score of 79.0%, an accuracy of 84.0%, a sensitivity of 73.0%, a specificity of 87.0%, a positive predicted value (PPV) of 64.0%, and a negative predicted value (NPV) of 92.0%. For the second model (severe *vs.* non-severe), an F1 score of 85.0%, a sensitivity of 82.0%, a specificity of 88.0%, a PPV of 77.0%, and an NPV of 91.0% was achieved. Together, the performance of both models combined was: an F1 score of 77.8%, a sensitivity of 76.8%, a specificity of 69.7%, a PPV of 79.8%, and an NPV of 85.9%. Based on the performance reported, this DMIND AI model is ready to be tested in real clinical settings with patients.

The 17-item Hamilton depression rating scale

This 17-item questionnaire is a widely recognized clinical tool for assessing depression. Each item scores depression symptoms on a scale from 0 to 4, with higher total scores indicating more severe depression. For this study, we used the Thai version that was culturally adapted by Manote L, *et al.* [11] to reflect local expressions of mood disorders. The kappa value of the HDRS-17 Thai was 0.87 with good internal consistency (standardized Cronbach's alpha coefficient = 0.7380).

Statistical analysis

Quantitative data were analyzed using appropriate statistical measures depending on the data distribution, including the mean, standard deviation (SD), median, and interquartile range (IQR). Gender, education level, and history of mental health conditions, were presented as frequencies and percentages to provide a comprehensive demographic profile of the participants. Our primary analysis was the extent of agreement in depression severity classifications between the DMIND application and HDRS-17 using Cohen's kappa coefficient with quadratic weighting. This measure was chosen as it is appropriate for ordinal data where the weight of disagreement varies.

Regarding the diagnostic performance of the DMIND application, various statistical metrics were

calculated: 1) accuracy - the proportion of true results; 2) F1 score - the harmonic mean of precision and recall; 3) sensitivity - the proportion of actual positive cases correctly identified by the model; 4) specificity - the proportion of actual negative cases correctly identified; 5) PPV/precision - the proportion of true positive results among all positive results predicted by the model; and 6) NPV - the proportion of true negative results among all negative results predicted by the model.

Results

Participant demographics

Our sample included 388 patients (93 males and 295 females) with a mean age of 37.2 ± 12.0 years and median age of 35 years. A majority of participants completed a Bachelor's Degree or higher (62.1%), whereas the remaining either had a secondary education (30.9%) or an elementary education (5.9%). Detailed demographic information is presented in **Table 1.** According to the HDRS-17, 269 patients (69.3%) were either normal or had mild depression, 70 patients (18.0%) were in the moderate group, and 49 patients (12.6%) were in the severe group.

Table 2 displays the agreement between the predicted depression classification by the DMIND AI model and the HDRS-17 across different severity levels (normal/mild, moderate, and severe). Each cell in the matrix contains the number and percentage of instances classified into each category. The results indicate substantial agreement between AI classification and HAMD-17 ratings for individuals with normal/mild depression at 91.4%. However, discrepancies are evident in the moderate depression category, where the AI's agreement with HAMD-17 is markedly lower at 27.6%. The AI tends to overclassify cases as moderate: 56.2% of those it classified as moderate were deemed normal/mild by HAMD-17, and 16.2% were deemed severe.

The DMIND AI model performance

Our results indicate that from the sensitivity, 87.3% of all positive cases identified by the model were indeed depressed (moderate/severe group), and 48.6% of all depressed individuals were correctly identified by the DMIND AI model, respectively. Looking at the specificity and NPV values, the DMIND AI model correctly identified 59.5% of all true non-depressed individuals, and 91.4% of all negative cases identified by the model were indeed non-depressed, respectively. Our DMIND application achieved an F1 score of 0.625 and had an accuracy of 0.68, which indicates it correctly identified 68.0% of all cases (positive and negative cases).

These results highlight the effectiveness of the DMIND application in accurately identifying depression cases. The relatively high values for sensitivity, NPV, and accuracy illustrate the potential of the DMIND application as a reliable tool for depression screening and monitoring.

Agreement between the HAMD-17 and the DMIND AI model using Cohen's Kappa [12]

To examine the agreement, we utilized the quadratic weighting method to reflect the proportional difference in the degree of disagreement among ratings. Findings in **Table 3** suggest there was consistent alignment across the two assessments.

Discussion

Our study introduces the complete DMIND application, comprised of a psychiatrist avatar that interviews the user and an AI depression scoring model. Here, we evaluate the AI model's performance and assess its agreement with a well-established depression rating scale.

The AI model demonstrated an 87.5% agreement with the HDRS-17 in the classification of depression severity, indicating a high level of concordance between the traditional assessment and the newly developed tool. Looking at the strength of this agreement using Cohen's kappa coefficient, a value of 0.45 was found, implying moderate agreement. Regarding diagnostic performance, the AI tool correctly identified 87.3% of individuals with depression (sensitivity), and correctly identified 59.5% of non-depressed individuals (specificity) and 91.6% of all negative cases identified by the model were indeed non-depressed. Taken together, the DMIND AI model demonstrated excellent performance metrics, highlighting its effectiveness in detect depression. Our results align with the findings of Milintsevich K, *et al.* [13] suggesting that their new model, a multi-target hierarchical regression model was trained on patient-psychiatrist interview transcripts from the DAIC-WOZ corpus designed to predict individual symptoms of depression, allowing for a fine-grained analysis of a patient's condition.

Table 1. Demographic characteristics.

Variables	Total (n = 388)	HAMD-17 (n = 388)			AI classification (n = 387)		
		Normal + mild depression (n = 269)	Moderate depression (n = 70)	Severe depression (n = 49)	Normal + mild depression (n = 175)	Moderate depression (n = 185)	Severe depression (n = 27)
Age (years)							
Mean ± SD	37.2 ± 12.0	39.0 ± 11.9	34.3 ± 12.1	31.8 ± 9.6	38.6 ± 11.3	36.3 ± 12.6	34.4 ± 11.2
Median (IQR)	35 (18)	38 (19)	30.5 (15)	30 (11)	37 (17)	33 (18)	30 (21)
Min-max	16 - 72	18 - 72	16 - 67	18 - 66	17 - 66	16 - 72	20 - 59
Gender, n (%)							
Female	295 (76.0)	197 (73.2)	56 (80.0)	42 (85.7)	119 (68.0)	151 (1.6)	24 (88.9)
Male	93 (24.0)	72 (26.8)	14 (20.0)	7 (14.3)	56 (32.0)	34 (18.4)	3 (11.1)
Religion							
Buddhist	352 (90.7)	256 (95.2)	60 (85.7)	36 (73.5)	166 (94.9)	164 (88.7)	22 (81.5)
Christian	6 (1.6)	4 (1.5)	1 (1.4)	1 (2.0)	2 (1.1)	3 (1.6)	1 (0.00)
Muslim	12 (3.1)	5 (1.9)	3 (4.3)	4 (8.2)	6 (3.4)	4 (2.2)	2 (7.4)
Others	18 (4.6)	4 (1.5)	6 (8.6)	8 (16.3)	1 (0.6)	14 (7.6)	3 (11.1)
Birthplace							
Central	262 (67.5)	185 (68.8)	44 (62.9)	33 (67.4)	117 (66.9)	126 (68.1)	18 (66.7)
North	28 (7.2)	16 (6.0)	7 (10.0)	5 (10.2)	13 (7.4)	14 (7.6)	3 (3.7)
Northeast	64 (16.5)	47 (17.5)	12 (17.1)	5 (10.2)	33 (18.9)	27 (14.6)	4 (14.8)
South	21 (5.4)	13 (4.8)	3 (4.3)	5 (10.2)	10 (5.7)	9 (4.9)	4 (7.4)
East	10 (2.6)	6 (2.2)	3 (4.3)	1 (2.0)	2 (1.1)	6 (3.2)	2 (7.4)
West	3 (0.8)	2 (0.7)	1 (1.4)	0 (0.0)	0 (0.0)	3 (1.6)	0 (0.00)
Marital status, n (%)							
Unmarried	229 (59.0)	146 (54.3)	46 (65.7)	37 (75.5)	99 (56.6)	111 (60.0)	18 (66.7)
Married	116 (29.9)	91 (33.8)	16 (22.9)	9 (18.4)	62 (35.4)	48 (26.0)	6 (22.2)
Divorced	42 (10.8)	31 (11.5)	8 (11.4)	3 (6.1)	13 (7.4)	26 (14.1)	3 (11.1)
Missing data	1 (0.3)	1 (0.4)	0 (0.0)	0 (0.0)	1 (0.6)	0 (0.0)	0 (0.00)
Education level, n (%)							
None	1 (0.3)	1 (0.4)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.5)	1 (0.0)
Elementary	23 (5.9)	15 (5.6)	5 (7.1)	3 (6.1)	8 (4.6)	13 (7.0)	2 (7.4)
Junior secondary	33 (8.5)	29 (10.8)	2 (2.9)	2 (4.1)	23 (13.1)	8 (4.3)	3 (7.4)
Senior secondary	87 (22.4)	62 (23.1)	13 (18.6)	12 (24.5)	40 (22.9)	40 (21.6)	7 (25.9)
Bachelor's Degree	206 (53.1)	135 (50.2)	42 (60.0)	29 (59.2)	87 (49.7)	103 (55.7)	15 (55.6)
Higher than Bachelor's Degree	35 (9.0)	25 (9.3)	7 (10.0)	3 (6.1)	16 (9.1)	18 (9.7)	1 (3.7)
Missing data	3 (0.8)	2 (0.7)	1 (1.4)	0 (0.0)	1 (0.6)	2 (1.1)	0 (0.0)
Occupation, n (%)							
Student	33 (8.5)	15 (5.6)	12 (17.1)	6 (12.2)	12 (6.9)	19 (10.3)	2 (7.4)
Government officer	148 (38.1)	141 (52.4)	4 (5.7)	3 (6.1)	103 (58.9)	41 (22.2)	4 (14.8)

Table 1. (Cont.) Demographic characteristics.

Variables	Total (n = 388)	HAMD-17 (n = 388)			AI classification (n = 387)		
		Normal + mild depression (n = 269)	Moderate depression (n = 70)	Severe depression (n = 49)	Normal + mild depression (n = 175)	Moderate depression (n = 185)	Severe depression (n = 27)
Contractor/ freelance	47 (12.1)	31 (11.5)	8 (11.4)	8 (16.3)	19 (10.9)	24 (13.0)	4 (14.8)
Employee of a private company	65 (16.8)	33 (12.3)	16 (22.9)	16 (32.7)	11 (6.3)	48 (26.0)	6 (22.2)
Self-owned business	25 (6.4)	10 (3.7)	10 (14.3)	5 (10.2)	3 (1.7)	19 (10.3)	3 (11.1)
Online merchants	2 (0.5)	1 (0.4)	1 (1.4)	0 (0.00)	1 (0.6)	1 (0.5)	0 (0.0)
Unemployed	45 (11.6)	17 (6.3)	18 (25.7)	10 (20.4)	9 (5.14)	27 (14.6)	7 (29.6)
Other	22 (5.7)	20 (7.4)	1 (1.4)	1 (2.0)	16 (9.1)	5 (3.2)	0 (0.0)
Missing data	1 (0.3)	1 (0.4)	0 (0.0)	0 (0.0)	1 (0.6)	0 (0.0)	0 (0.0)
Income, n (%) Baht							
< 10,000	47 (12.1)	34 (12.6)	9 (12.9)	4 (8.2)	20 (11.4)	22 (11.9)	5 (18.5)
10,000 - 20,000	107 (27.6)	79 (29.4)	15 (21.4)	13 (26.5)	60 (34.3)	38 (20.6)	9 (33.3)
> 20,000 - 30,000	88 (22.7)	63 (23.4)	13 (18.6)	12 (24.5)	39 (22.3)	45 (24.3)	4 (14.8)
> 30,000 - 40,000	114 (29.4)	77 (28.6)	21 (30.0)	16 (32.7)	49 (28.0)	57 (30.8)	8 (29.6)
> 40,000	23 (5.9)	12 (4.5)	8 (11.4)	3 (6.1)	5 (2.9)	17 (9.2)	0 (0.0)
Missing data	9 (2.3)	4 (1.5)	4 (5.7)	1 (2.0)	2 (1.1)	6 (3.2)	1 (3.7)
Medical history, n (%)							
No	275 (70.9)	194 (72.1)	47 (67.1)	34 (69.4)	131 (74.9)	124 (67.0)	19 (70.4)
Yes	113 (29.1)	75 (27.9)	23 (32.9)	15 (30.6)	44 (25.1)	66 (33.0)	8 (29.6)
NCD (noncommunicable diseases)	72 (63.7)	47 (62.7)	19 (82.6)	6 (40.0)	25 (56.8)	41 (67.2)	6 (75.0)
Neurology	8 (7.1)	5 (6.7)	2 (8.7)	1 (6.7)	3 (6.8)	5 (8.2)	1 (0.0)
Endocrine	20 (17.7)	16 (21.3)	1 (4.4)	3 (20.0)	12 (27.3)	5 (11.5)	2 (12.5)
Gastrointestinal	9 (8.0)	21 (8.0)	1 (4.4)	2 (13.3)	3 (6.8)	4 (6.6)	3 (25.0)
Gynecology	7 (6.2)	4 (5.3)	1 (4.4)	2 (13.3)	1 (2.3)	6 (9.8)	0 (0.0)
Psychiatric history, n (%)							
No	192 (49.5)	185 (68.8)	3 (4.3)	4 (8.2)	135 (77.1)	49 (26.5)	8 (29.6)
Yes	196 (50.5)	84 (31.2)	67 (95.7)	45 (91.8)	40 (22.9)	136 (73.5)	19 (70.4)
Substance use history, n (%)							
No	351 (90.5)	258 (95.9)	56 (80.0)	37 (75.5)	169 (96.6)	160 (86.5)	21 (77.8)
Yes	37 (9.5)	11 (4.1)	14 (20.0)	12 (24.5)	6 (3.4)	25 (13.5)	6 (22.2)
Family history psychiatric disorder, n (%)							
No	337 (86.9)	241 (89.6)	57 (81.4)	39 (79.6)	164 (93.7)	149 (80.5)	24 (88.9)
Yes	51 (13.1)	28 (10.4)	13 (18.6)	10 (20.4)	11 (6.3)	36 (19.5)	3 (11.1)

HDRS-17, 17-item Hamilton depression rating scale; DMIND AI Model, Detection and monitoring intelligence network for Depression AI model; SD, Standard deviation; IQR, Interquartile range.

Table 2. The performance of the DMIND application compared to the HDRS-17 classifications.

Factor	AI Classification		
	Normal/ mild	**Moderate**	**Severe**
HAMD-17			
Normal/mild	160 (91.4)	104 (56.2)	5 (18.5)
Moderate	10 (5.7)	51 (27.6)	9 (33.3)
Severe	5 (2.9)	30 (16.2)	13 (48.2)

Results are displayed as the number (percentage) of instances classified into each category

Table 3. Agreement between the HAMD-17 and the DMIND Al model using Cohen Kappa.

Pairs	Agreement	*Kappa*	Interpretation
HAMD-17 and DMIND AI	87.5%	0.45	Moderate

This approach shifts attention from a binary classification of depression to a personalized analysis of symptom profiles achieving depression classification (F1 = 73.9). Bauer B, *et al.* [14] supported the performance of AI in detecting suicidal ideation by utilizing large language models (LLMs) to analyze Reddit discussions on suicidality, specifically focusing on the r/SuicideWatch subreddit. Their study examined 2.9 million posts from 30 subreddits and identified that posts in r/SuicideWatch expressed feelings of disconnection, burdensomeness, hopelessness, and trauma. These findings align with the Hierarchical taxonomy of psychopathology (HiTOP), which includes elements like seeking support and the severity of distress. This research confirms that language-based theories of suicide are valid and demonstrates the potential of natural language processing (NLP) in understanding online emotional expressions and validating mental health theories.

The high sensitivity of the DMIND AI is crucial for the rapid identification of potential depression cases, allowing for swift therapeutic intervention. The high NPV further enhances its reliability as a screening tool, ensuring that individuals without depression are accurately identified and reducing unnecessary concerns. While the low specificity and low PPV might initially seem like drawbacks, they are less critical in the context of a screening tool. The primary goal is to avoid missing any cases, especially those at risk of suicide. Following an initial AI-based screening, flagged individuals can undergo further assessment by a psychiatrist. This approach ensures a broader net is cast, capturing all potential cases and prioritizing safety over the risk of false positives, which can be addressed through subsequent professional evaluation.

Findings of a substantial agreement between the DMIND and HDRS-17 validate the DMIND application as a credible alternative to human assessments, especially in environments with limited access to mental health professionals.

The AI-incorporated DMIND application is a promising tool for routine depression screening in various healthcare settings as well as through online channels. Generally, hospitals tend to rely on the paper questionnaires/interviews for screening patients. The use of AI tools will increase screening accuracy through its capability to detect patients' emotions from their responses. Screening through avatar instructions in the DMIND application would help users/patients feel more comfortable in expressing their feelings and opinions, allowing healthcare staff to gain more in-depth information.

The DMIND application is anticipated to transform the management and treatment of depression in Thailand. In line with Squires M, *et al.* [15], these data-driven methods promise more precise and personalized approaches to 1) detecting, diagnosing, and treating depression; 2) identifying individuals with mental health conditions; and 3) tailoring treatments to those who will benefit most. Additionally, unsupervised learning techniques are revealing the extensive heterogeneity within depression diagnoses, moving away from traditional discrete diagnostic categories and toward evidence-based treatment. With AI and big data, this application helps take us one step closer to depression prevention and treatment in Thailand.

The strengths of this study lie in its innovative integration of self-assessment tools, high-quality validation against an established measure, expert-driven design, emotionally engaging avatar

interaction, and robust behavioral data analysis.

The DMIND application is one of the first applications in Thailand to integrate advanced AI technologies with a screening assessment to enhance diagnostic precision and efficiency by analyzing behavioral data. The underlying AI model demonstrated excellent validity metrics, highlighting its strength in identifying true depressed and non-depressed patients.

Additionally, the user interactions with the psychiatrist avatar were designed to make patients feel understood and supported, differentiating it from traditional chatbots, and helps create a comfortable environment for patients to express their emotions freely. Throughout the data collection process, it was observed that many patients cried and talked extensively with the avatar. Patients provided feedback saying that they felt relaxed and comfortable sharing their feelings with the avatar, illustrating the unique strength of the application in fostering emotional engagement and providing a therapeutic experience.

Moreover, following a structured protocol period, the study leveraged behavioral data to capture emotions via text. This comprehensive assessment of depression, analyzing both the content and sequence of responses, was developed based on an extensive literature review and consultations with renowned psychiatrists.

Another notable strength is that the DMIND application can be accessed and used on most digital devices. This flexibility offers a non-intrusive, reachable channel for the early detection and monitoring of depression, which is particularly advantageous in cultural contexts where mental health issues may be stigmatized or in rural areas where transportation is scarce.

Whilst the study presents promising results, several limitations must be acknowledged. The sample size, although sufficient for preliminary analysis, does not represent the broader population. Our sample only included patients recruited from hospitals, thus, individuals from other backgrounds or those who may experience milder forms of depression have been underrepresented. Moreover, the exclusion of people with severe psychiatric conditions or cognitive impairments may skew the applicability of the DMIND tool as these complex conditions are also associated with depression. The use of purposive sampling in our study introduces certain limitations. Since our sample is not randomly selected, it might not accurately represent the general population, which can restrict the findings' generalizability. There's also a higher risk of selection bias, as the sample might mirror the researchers' subjective choices more than the broad population characteristics. These limitations could influence how the study results are interpreted and the DMIND application's applicability across various settings or populations

One of the most significant limitations that may have impacted the agreement between our model and the HDRS-17 was that a trained psychiatrist did the data labelling for the AI model used this study. Accordingly, the labels for the AI model may have aligned more closely with the DSM-5 criteria rather than the HDRS-17. As for the limitations that show moderate specificity, they may arise from the AI model's design, which favors sensitivity to ensure that nearly all cases of depression are detected, even if it means higher false positives. Setting the diagnostic threshold low enhances sensitivity at the expense of specificity, leading to more healthy individuals being incorrectly classified as depressed. This trade-off is common in mental health diagnostics where missing true cases can have more severe consequences than false positives. The model prioritizes capturing all potential instances of depression, considering the significant impacts of undiagnosed depression.

Another significant limitation of the DMIND application is its reliance on digital technology. Older individuals or those less familiar with digital devices may face challenges in using the application, potentially limiting its accessibility for these populations. Despite being trained on a large dataset, the model may still contain cultural biases, as its current algorithm and training data may not fully capture all cultural variations in expressing and interpreting symptoms of depression. Consequently, minor or uncommon signs of depression might have been overlooked during model training.

Additionally, the model currently struggles to classify depression into four distinct categories (normal, mild, moderate, and severe), which would be more beneficial for identifying and managing patients. This limitation is due to the insufficient number of severe cases available for training the AI model, preventing the achievement of a four-level classification in this study.

Even though the current study compared the performance of the DMIND application with the HDRS-17, this scale's accuracy in identifying

depression severity is still limited. To address this, further research should use expert psychiatrists as the gold standard in screening patients for depression instead of using the HDRS-17. As the past video interviews have been stored online with consent from participants, we aim to have psychiatrists review all video clips, determine the participant's depression severity, and use this as the true diagnosis instead. By comparing the AI predictions from the DMIND application to both the HDRS-17 and true diagnosis from experts, a more robust evaluation of the DMIND model's performance can be done. This approach could lead to better model accuracy and reliability.

Furthermore, future studies could also compare the AI model predictions with other depression screening tools to confirm the DMIND application's agreement across a larger range of assessments and ensure its applicability as an alternative depression screening tool on a wider scale. Concurrently, further studies should investigate the effectiveness of alternative AI algorithms and multimodal models that analyze other depression-related features in the context of model performance and user acceptance.

Notably, the prospect of a simplified version of the DMIND questionnaire should be examined for easier pre-screening. A reduced set of questions could be integrated with the interactive voice response (IVR) system of the hotline service to allow depression screening via phone calls.

Exploration of the integration of the DMIND system with other technologies, such as wearable devices, could lead to the possibility of real-time data collection and ongoing depression management, paving the way to more personalized and dynamic treatment plans. Long-term studies would be valuable to assess the effectiveness of DMIND in monitoring depression progression and its response to treatment over extended periods.

We will continue to use comprehensive performance metrics to periodically assess the model's performance and implement necessary optimizations. In the near future, we plan to collect more data to improve the model's generalizability and robustness and expand the dataset to include diverse demographic groups and various clinical settings. With this, the AI could be updated to recognize and interpret a wider range of depressive expressions in a larger variety of cultures. Simultaneously, patient privacy and data security will remain paramount, with all data handling and processing adhering to strict ethical guidelines and regulations.

Finally, In address the AI intervention's potential misuses, it is crucial to caution against replacing professional medical judgments with AI outputs, and the risks of over-diagnosis and data privacy breaches. Future research should develop strict usage guidelines, improve algorithm accuracy, and ensure AI integration respects patient confidentiality and involves clinician oversight.

Conclusion

Our study provides evidence supporting the use of the DMIND AI-assisted application in diagnosing and assessing the severity of depression. The moderate agreement between DMIND and HDRS-17, along with its strong sensitivity, underscores its potential as a scalable and accessible solution for depression management. This application can serve as a critical instrument in both psychiatric and primary care settings, aiding clinicians in the rapid, reliable, and early assessment of depression. For that reason, the DMIND application, and other similar tools, will be particularly helpful in areas where access to mental health services and staff is limited. Future research and implementation efforts should focus on further validating these findings and exploring the application's integration into clinical practice to enhance mental health care delivery.

Acknowledgements

The authors would like to express deep gratitude to all of the subjects who were involved in this study.

Conflicts of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported.

Data sharing statement

All data generated or analyzed during the present study are included in this published article. Further details are available for noncommercial purposes from the corresponding author on reasonable request.

References

1. Prasartpornsirichoke J, Pityaratstian N, Poolvoralaks C, Sirinimnualkul N, Ormtavesub T, Hiranwattana N, et al. The prevalence and economic burden of treatment-resistant depression in Thailand. BMC Public Health 2023;23:1541.

2. Nochaiwong S, Ruengorn C, Thavorn K, Hutton B, Awiphan R, Phosuya C, et al. Global prevalence of mental health issues among the general population during the coronavirus disease 2019 pandemic: a systematic review and meta-analysis. Sci Rep 2021;11:10173.
3. Department of Mental Health's recent surveys and studies December 2022.
4. Moreno MA, Jelenchick LA, Egan KG, Cox E, Young H, Gannon KE, et al. Feeling bad on Facebook: depression disclosures by college students on a social networking site. Depress Anxiety 2011;28: 447-55.
5. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. Proceedings of the International AAAI Conference on Web and Social Media 2021;7:128-37.
6. Cummins N, Dineley J, Conde P, Matcham F, Siddi S, Lamers F, et al. Multilingual markers of depression in remotely collected speech samples: a preliminary analysis. J Affect Disord 2023;341:128-36.
7. Hemrungrojn S, Saengsai K, Jakkrawankul P , Kiattiporn-Opas C, Chaichareenon K, Amrapala A, et al. Development and evaluation of the DMIND questionnaire: Preparing for AI integration into an effective depression screening tool. medRxiv 2024: 2024.06.07.24308625.
8. Obuchowski NA. Sample size calculations in studies of test accuracy. Stat Methods Med Res 1998;7: 371-92.
9. Serengil SI, Ozpinar A, editors. LightFace: A hybrid deep face recognition framework. 2020 Innovations in intelligent systems and applications conference (ASYU); 2020 Oct 15-17; Istanbul, Turkey: IEEE; 2020.
10. Zhang X, Malkov Y, Florez O, Park S, McWilliams B, Han J, et al, editors. TwHIN-BERT: a socially-enriched pre-trained language model for multilingual tweet representations at twitter. 29th ACM SIGKDD conference on knowledge discovery and data mining, KDD 2023; 2023 Aug 6; Long beach, United States: Association for computing machinery; 2023.
11. Lotraku M, Sukanich P, Sukying C. The reliability and validity of Thai version of Hamilton rating scale for depression. J Psychiatry Assoc Thailand 1996;41; 235-246.
12. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977;33:363-74.
13. Milintsevich K, Sirts K, Dias G. Towards automatic text-based estimation of depression through symptom prediction. Brain Inform 2023;10:4.
14. Bauer B, Norel R, Leow A, Rached ZA, Wen B, Cecchi G. Using large language models to understand suicidality in a social media-based taxonomy of mental health disorders: linguistic analysis of Reddit posts. JMIR Ment Health 2024;11:e57234.
15. Squires M, Tao X, Elangovan S, Gururajan R, Zhou X, Acharya UR, et al. Deep learning and machine learn ing in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. Brain Inform 2023;10:10.