An Application for Mental Health Monitoring Using Facial, Voice, and Questionnaire Information

Suphalerk Boonvitchaikul 1, Napat Cheetanom 1 *, Tagon Sompong 1 *, Jirapat Sununtnasuk 1, Siri Thammarerkrit 1, Pattaraporn Pongpanatatipat 1, Punnaphoj Aeuepalisa 2, Ananya Kuasakunrugroj 2, Chatavut Viriyasuthee 2, Patawee Prakrankamanant 1, Sorawit Wainipitapong 3, Ekapol Chuangsuwanich 1

1Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand
2Vulcan Coalition Co., Ltd, Bangkok, Thailand
3Department of Psychiatry, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

{6230524921, 6231317521, 6232006521, 6230067421, 6232029021, 6572059621}@student.chula.ac.th, {punnaphoj, ananya.k, chatavut}@lab.ai, 6170204021@alumni.chula.ac.th, sorawit@chula.md, ekapolc@cp.eng.chula.ac.th

Abstract
Depression is a major societal issue. However, depression can be hard to self-diagnose, and people suffering from depression often hesitate to consult with professionals. We discuss the design and initial testing of our prototype application that performs depression detection using multi-modal information such as questionnaires, speech, and face landmarks. The application has an animated avatar ask questions concerning the users’ well-being. To perform screening, we opt for a 2-stage method which first predicts individual HAM-D ratings for better explainability, which may help facilitate the referral process to medical professionals if required. Initial results show that our system archives 0.85 Marco-F1 for the depression detection task.

Introduction
Depression is a mood disorder that affects a person’s thoughts, feelings, and behavior. An increasing number of people are experiencing problems with depression compared to the past (Organization et al. 2017). If depression is detected early, it can be promptly treated. However, people are unaware of the risk of depression which leads to some cumulative stress. Some might avoid seeing out or disclosing their problems to a psychiatrist. Additionally, the number of psychiatrists available is insufficient for the number of patients who require consultation (Butryn et al. 2017; Kakuma et al. 2011), leading to long waiting times, which can cause delays in initiating treatment. A simple to use screening software application might be able to help reduce the availability problem and can help lead more people with high depressive symptoms to seek professional help (BinDhim et al. 2016).

Several works have explored the use of automatic depression screening using machine learning techniques. Text and video/audio information were often used as the main components. Examples of text-based screening include using social media text (Jain et al. 2019; Lin et al. 2020) or transcribed interviews (Devlin et al. 2018; Senn et al. 2022). Similarly, when using video/audio information, basic features such as sound quality and spectral analysis were often used (Yalamanchili et al. 2020), and pre-trained models such as Face Action Units (FAU) (Williamson et al. 2016; Valstar et al. 2016) or emotional classifier models are used to extract audio information (Williamson et al. 2016).

For software applications, questionnaire-based screening methods are often used (BinDhim et al. 2015; Yalamanchili et al. 2020; Ziwei and Chua 2019). Although these provide effective screening measures, self-administered questionnaires might be less accurate than professionally conducted interviews since the interviewee can provide more context (Guohou, Lina, and Dongsong 2020).

In this work, we aim to create an application for depression screening that has the following desirable properties: 1) user-friendly 2) effective in utilizing the rich information from voice, facial, and questionnaire information 3) supporting the delivery of clear and easily transferable critical patient information to psychiatrists upon their request. To this end, we create a prototype application that includes a virtual assistant that can provide basic consultations to patients. Our system incorporates avatars that ask questions and collect user interaction data via questionnaires, speech, and facial expressions of users which are then used to analyze the depression risk. Unlike previous works in machine-learning-based depression screening, we design the system such that the depression risk is broken down into different factors which can help the doctor quickly assess and continue from the information gathered by the application. This work presents the early results of our application in terms of its predictive capabilities.

Depression Screening and Detection
Depression detection can be categorized into 2 main types: self-administration and clinician-administration. Self-administration depression detection does not require a
professional or clinician, such as Beck Depression Index (Jackson-Koku 2016) and Quick Inventory of Depressive Symptomatology (QIDS) (Rush et al. 2003). On the contrary, Hamilton Depression Rating Scale (HAM-D) (Hamilton 1960) and Montgomery-Asberg Depression Rating Scale (MARS-D) (Davidson et al. 1986) are clinician-administered. The clinician conduct interviews and notes the condition of the interviewee accordingly. Clinician-administered methods are considered more effective than self-administered ones (Guohou, Lina, and Dong-song 2020). First, the interviewee can provide detailed and unstructured responses to each question, which enables the clinician to make a more accurate assessment. Second, a clinician is able to personalize questions based on the previous session. Thus, a depression screening method that can utilize rich and diverse sets of questions and their responses should be able to provide better detection performance. The HAM-D and MARS-D have been shown to have a high correlation between the two scores. We chose HAM-D as our screening standard because it also includes several parts that are more physically apparent, such as slowness in speech, which should be more observable by automated techniques. The HAM-D consists of 17 items of depression symptoms, each representing different aspects such as HAM-D3 is about the presence of suicidal thoughts. Each symptom is graded to its corresponding scale and summed to obtain the final HAM-D score.

Data Collection
We used DAIC-WOZ (Gratch et al. 2014) as a guideline in our data collection where medical professionals were involved in analyzing the results. We recruited participants in Thailand to use our application in a controlled setting. A medical professional is observing each participant as he/she uses the application and rates the participant using the HAM-D guidelines. Our mobile application interface is shown in Fig. 1. There are two kinds of questions namely, questionnaire questions and freeform questions. The questionnaire questions first require the users to answer questions with multiple choices or checkboxes. Then, they are asked to talk about their answers. The speech and facial expressions of the users are recorded via the application. Rather than recording the video of the face, to protect the user's identity we opt to process the video on device and only record the facial landmarks. The freeform questions only have the user answer the question by talking to the application. The distribution of our data is shown in Table 1. Note that only a quarter of the users opt to turn on the camera. Only audio recordings are available for the rest of the users.

System Design

System Overview
Our system takes in the data from different modalities in order to assess the depression risk of the user as shown in Fig. 2. To provide more explainability, we opt for a two-stage approach where the 17 questions of the HAM-D are predicted first. Then, a second-stage model is used to perform the final depression risk score. The speech information is transcribed using an Automatic Speech Recognizer (ASR) to provide the transcription for textual analysis. Speech is utilized to capture any information that might be presented in the prosody and intonation. We set “no symptom” as the negative class and others as the positive class in any particular HAM-D task. The system consists of 4 kinds of models for each modality: facial model, text model, audio model, and questionnaire model. Specific HAM-D items are predicted using information from the questions related to the particular item. In the second stage, we use XGBoost’s regression model (Chen et al. 2015) to aggregate the results as a depression risk score.

Questionnaire Model
The questionnaire model is an XGBoost’s regressor. Questionnaire answers were treated as tabular data by concatenat-
Figure 2: The system’s pipeline involves gathering user data, which includes facial videos, speech, and questionnaire answers. Each type of the information is then sent to its corresponding model to predict each HAM-D individually. A final aggregate model classifies whether a person is at high-risk or not.

ing all one-hot vectors of the questionnaire answers. In general, these models are very strong for most HAM-D items. For most HAM-D items, we only use the outputs from the questionnaire model.

Text Model
We used WangchanBERTa, a RoBERTa-based pre-trained model for the Thai language, to perform text classification. Furthermore, we also used an ensemble method which fused the output from text and questionnaire models.

Facial Model
The facial model aims to improve HAM-D8 (retardation diagnosis) which is related to motion and expression. We create features called “Motion Vector” as follows:

\[ \overrightarrow{M} = \frac{1}{T - 1} \sum_{n=1}^{T-1} [v_{1,t} \ldots v_{N,t}]^\top \]
\[ v_{i,t} = \frac{\|\overrightarrow{p}_{i,t} - \overrightarrow{p}_{i,t+1}\|_2}{\|\overrightarrow{p}_{i,t} - \overrightarrow{p}_{i,t+1}\|_2} \]
\[ \overrightarrow{p}_{i,t} = [x_{i,t}, y_{i,t}, z_{i,t}]^\top \] (1)

where \( T \) is the number of frames, and \( N \) is the number of vertices in a facial landmark. \( \overrightarrow{p}_{i,t} \) denotes the coordinates of facial landmark \( i \) at time \( t \). A multi-layer perceptron model is chosen for the model. Besides motion, questionnaire and Word Per Second (WPS), features can also be included.

Table 2: F1 of different models for certain HAM-D items. * indicates a statistically significant difference (paired t-test) compared to the questionnaire model (\( p < 0.05 \)).

<table>
<thead>
<tr>
<th>Item</th>
<th>Best Model</th>
<th>Best</th>
<th>Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Text &amp; Questionnaire</td>
<td>0.80</td>
<td>0.69</td>
</tr>
<tr>
<td>6</td>
<td>Text</td>
<td>0.61</td>
<td>0.43</td>
</tr>
<tr>
<td>10</td>
<td>Text</td>
<td>0.75*</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table 3: Results on HAM-D8 with different modalities.

<table>
<thead>
<tr>
<th>Features</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questionnaire</td>
<td>0.32</td>
<td>0.37</td>
<td>0.34</td>
</tr>
<tr>
<td>Motion</td>
<td>0.61</td>
<td>0.63</td>
<td>0.60</td>
</tr>
<tr>
<td>Questionnaire + Motion</td>
<td>0.77</td>
<td>0.82</td>
<td>0.79</td>
</tr>
<tr>
<td>Questionnaire + Motion + WPS</td>
<td>0.82</td>
<td>0.85</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Table 2: F1 of different models for certain HAM-D items. * indicates a statistically significant difference (paired t-test) compared to the questionnaire model (\( p < 0.05 \)).

Text Model
Particular HAM-D items that would benefit from textual analysis were chosen to be used with the text model. These included items that have questionnaires which are subjective from the participant’s point of view such as HAM-D1 (Depressed Mood), HAM-D6 (Early Hours Of The Morning), and HAM-D10 (Anxiety Psychic). We perform the data into 3 folds for training and evaluation and report the average across the folds. The best-performing models and their questionnaire-only counterparts are summarized in Table 2.

Facial Model
In our evaluation of the HAM-D8 binary classification task, we adopted the Leave-One-Out (LOO) method because there are only 21 samples for this experiment. As shown in Table 3, the model with multiple modalities (questionnaire, motion, and WPS) significantly outperforms the questionnaire-only model using the McNemar test with \( p < 0.05 \).

Fig. 3 illustrates how the facial landmark features differ between negative and positive classes. The biggest differences correspond to the landmarks at the chin (landmark 7-10) and the mouth (landmark 54-60).
Figure 3: The dissimilarity in facial features between negative and positive classes. The indices correspond to different landmark location.

<table>
<thead>
<tr>
<th>Individual HAM-D items</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questionnaire model</td>
<td>0.75</td>
<td>0.79</td>
<td>0.75</td>
</tr>
<tr>
<td>Best-multimodal model</td>
<td>0.85</td>
<td>0.89</td>
<td>0.85</td>
</tr>
<tr>
<td>Actual</td>
<td>0.94</td>
<td>0.96</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table 4: Comparison between using scores of intermediate HAM-D items from different sources.

**Aggregate Model**

We evaluate the performance of our aggregation model using a similar setting to the facial model where a LOO method is used on the 21 participants with facial features. The results are summarized in Table 4. Using HAM-D items from the questionnaire models yields worse performance than using the scores from the multimodal models. However, it is still worse than the ideal scenario where the actual rating given by the professional is used.

**Conclusion**

We describe our prototype application for depression detection which can utilize multi-modal inputs namely, questionnaire responses, facial landmarks, speech, and transcribed text. Early results suggest that our system can help depression screening archiving an F1 of 0.85. The system can also provide transcripts and assessments for each HAM-D item which can be used by doctors for further examination.

**Acknowledgements**

This work used high performance computing resources of the Center for AI in Medicine (CU-AIM), Faculty of Medicine, Chulalongkorn University, Thailand.

**References**


