Detecting Distress Changes Using Multimodal Data During Interaction with A Smart Speaker

Chingyuan Lin[†], Yuki Matsuda[‡,†], Hirohiko Suwa[†], Keiichi Yasumoto[†]

[†] *Nara Institute of Science and Technology*, Nara, Japan
{lin.chingyuan.la8, h-suwa, yasumoto}@is.naist.jp

[‡] *Okayama University*, Okayama, Japan
yukimat@okayama-u.ac.jp

***Abstract*—Mental health has a huge impact on humans, affecting both psychological and physical well-being. Excessive stress can lead to depression, reduced productivity, and even suicidal tendencies. Stress also impacts appetite and sleep quality, potentially leading to other health issues. However, stress accumulation often goes unnoticed until it severely impacts health, highlighting the need for daily stress level assessment. This study aims to estimate daily distress levels through natural conversations with a smart speaker. We utilize the audio-visual data of users interacting with a smart speaker on a daily basis, extract features from different modalities through analysis, and predict distress changes in daily life using questionnaire responses as labels. In the experiment, participants interacted with a smart speaker placed in their bedrooms, simulating daily life. Webcam recordings captured facial expressions, voice, and heart rate data, which were preprocessed for analysis. Predictions for happiness, depression, and anxiety levels were made using data from questionnaires filled out after each recording session, with scores ranging from 0 to 18. Results from the 14-day experiment with seven participants, aged 22 to 24, revealed MAEs of 2.04, 2.59, and 2.31 for happiness, depression, and anxiety levels, respectively. The corresponding RMSEs were 2.63, 3.20, and 2.91.**

***Index Terms*—distress, happiness, anxiety, depression, audio-visual, heart rate, multimodality, smart speaker**

I. Introduction

In recent years, there has been a significant shift in workplaces towards remote work, with approximately 12.7% of full-time employees working from home and 28.2% adopting hybrid models by 2023. In Tokyo, Japan, over half of businesses have embraced remote work, continuing for three consecutive years, with rates over 50% by the end of 2022. While indoor work environments reduce physical health risks, concerns about psychological well-being, including stress and depression, have grown. Unlike physical ailments, psychological distress develops gradually and is challenging to detect early on. Early prevention and intervention are crucial, this study aims to track changes in users' daily indicators of emotional distress, thus helping them keep abreast of their mental state. We intend to utilize facial expressions, voice, and heart rate data to measure the intensity of happiness, depression, and anxiety experienced by participants. In terms of data collection, the experimental setup is built in the participants' home to ensure their comfort. In the context of contemporary living, smart speakers have become ubiquitous in households, providing essential daily information and proving invaluable for health. The interactions with these devices, initiated by the users and typically brief, are adaptable across various age groups. Therefore, we used the smart speaker as a key tool in the experimental design, utilizing it to capture and analyze communication patterns with participants.

Many studies have analyzed distress levels using open datasets [1], but most of these datasets do not conform to realistic scenarios. In this experiment, we collected video data of 7 participants interacting with a smart speaker over 14 days. From these video data, we extracted three types of features: facial expressions, voice, and heart rate. We then used Random Forest Regression Model and LightGBM Regression Model to predict the changes in the levels of Happiness, Depression, and Anxiety. The label values were derived from the Depression and Anxiety Mood Scale (DAMS) questionnaires, which participants filled out after each recording session. Each emotion has a label value ranging from 0 to 18. The best MAE from leave-one-day-out cross-validation achieved for Happiness, Depression, and Anxiety were 2.04, 2.59, and 2.31, respectively, while the corresponding RMSE values were 2.63, 3.20, and 2.91. With this system, people do not need to collect data in unfamiliar environments or wear cumbersome sensors deliberately. They can obtain their daily distress variations in their familiar homes in a manner that conforms to realistic scenarios.

II. Related Work

Many studies have used different methods to improve accuracy in predicting the categories of emotions such as stress, anxiety, and happiness. Considering the following experiment procedure and model training, we categorize these diverse studies based on three distinct focal points for discussion. In Section II-A and Section II-B, we list some of the studies focusing on the utilization of audio-visual data to predict emotions. Moving on to Section II-C, we explore how previous studies obtained physiological signal information and how to utilize its representations. In Section II-D, we summarize the problems in these related works and explore suitable research approaches.

A. Emotion Recognition Using Visual Data

Facial cues are frequently employed to assess emotion and stress levels, given their capacity to communicate subtle non-verbal signals. The scrutiny of visual data enriches emotional evaluation, bolstering its thoroughness and precision, thus serving as a crucial asset in psychological investigations.

Soleymani *et al.* utilized the MAHNOB-HCI database, employed face tracker technology to detect landmarks from features, and achieved the highest accuracy on the LSTM model [2]. In another work, Duncan *et al.* used their custom-trained VGG-S network with a face-detector and Haar-Cascade filter provided by OpenCV to implement real-time facial emotion recognition [3]. They choose the extended Cohn-Kanade dataset (CK+) [4] and Japanese Female Facial Expression (JAFFE) database [5], analyzing the six common expressions (anger, fear, neutral, happy, sad, surprise). The overall training accuracy and test accuracy are 90.9% and 57.1%, respectively.

B. Emotion Recognition Using Audio Data

Audio, much like facial expressions, plays a crucial role in emotion and stress level analysis due to its capacity to convey subtle verbal and non-verbal nuances. For example, Spectrogram [6] visually represents the timbre, pitch, and rhythm of a voice and transforms audio signals into visual data, revealing patterns and intricacies in spoken language that are key to identifying emotional states.

C. Emotion Recognition Using Physiological Signal

Alongside facial expressions and audio signals, physiological indicators such as heart rate variability (HRV) and electrocardiogram (ECG) patterns play a vital role in forecasting emotional categories or indices. These metrics offer insights into the body's reactions to different emotional states and their impact on overall health and wellness.

To analyze physiological signals obtained from various devices or sensors, Santamaria-Granados *et al.* employed a deep learning approach using a DCNN [7]. Their study focused on a dataset of physiological signals, specifically the AMIGOS dataset. The detection of emotions in their study is achieved by correlating these physiological signals with the arousal and valence data from this dataset, aiming to classify the affective state of a person accurately. Their model demonstrated an impressive accuracy of 0.76 for Arousal, outperforming other models such as MESAE, traditional DNN, and CNN.

D. Research Directions

Most of the studies mentioned earlier focus on emotions indicators using audio-visual data and datasets produced in controlled environments. Participants usually stay in a specified room where they watch certain videos or read scripted speeches. However, this intentional setup for experiments may not be aligned with realistic scenarios, as it might not represent real-life situations accurately. Therefore, the following items are the key points that we would like to emphasize in the experiment of this study:

- Simple experimental conditions
- A familiar activity space for participants
- Short-time data collection
- Aligning with a realistic scenario

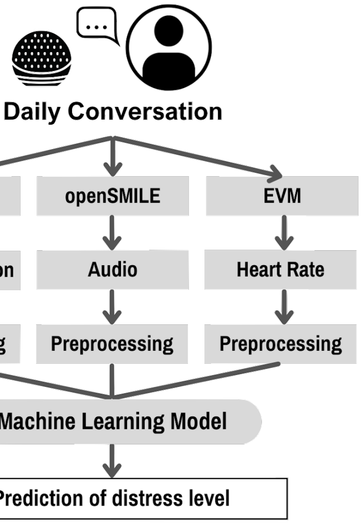

Fig. 1: Process of Data Analysis

Regarding the solution of the items shown above, we describe it in more detail in the proposed method.

III. Proposed Method

The purpose of this study is to enable individuals to monitor changes in their distress levels. To achieve this, we intend to create a system capable of collecting audio-visual data from interactions between individuals and smart speakers in their daily lives. The flowchart in Fig. 1 illustrates the process, with the environment setup as well as data collection at the top, data processing and feature extraction in the middle, and computation of results using various machine learning models at the bottom. We will explain each step in the following subsections.

A. Assumed Environment and Data Collection

In this study, unlike most related studies, we chose a smart speaker as the device to communicate with participants. We set the recording location in the participants' private rooms, mainly because participants could express their emotions more authentically at home. On the other hand, setting the experiment location in the participants' rooms reduces the complexity of the experiment. In addition, smart speakers can also provide essential daily information and hold significant purposes in the domains of health and education [8]. The interactions with these devices, initiated by the users and typically brief, are adaptable across various age groups [9]. A web camera was installed near the smart speaker to capture the participants' interactions, with each session lasting 40 seconds to cover around two question-answer exchanges.

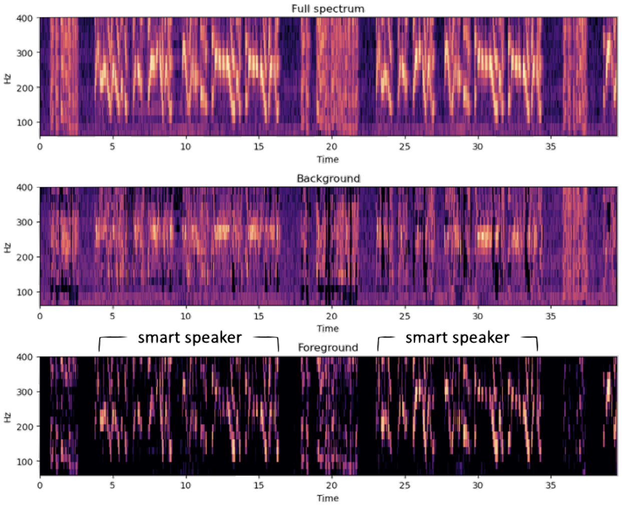

Fig. 2: Vocal Separation

B. Feature Extraction

In this subsection, we discuss the extraction of features for different modal types (face expression, audio, heart rate) respectively.

1) Face Expression Feature: For facial expressions, we utilized the OpenFace toolkit [10], an open-source tool for facial behavior analysis presented by Tadas Baltrusaitis *et al.*. It is useful for computer vision, machine learning, and affective computing. OpenFace specializes in facial landmark detection, head pose estimation, facial action unit recognition, and eye gaze estimation. This study focused on extracting six eye gaze and 23 facial action unit features, such as inner eyebrow raise and blinking, from a 40-second video clip. In addition, we calculated the mean and standard deviation of these 29 features.

2) Audio Feature: For audio features in video clips, we utilized OpenSMILE, an open-source tool proposed by Florian Eyben *et al.*, for extraction. We chose as a feature set the extended Geneva Minimal Acoustic Parameter Set (eGeMAPS) [11], which combines spectral parameters, temporal features, frequency-dependent parameters, and more. It contains a total of 88 parameters.

Because the webcam captures both the participant's voice and that of the smart speaker during recordings, our study specifically requires the participant's voice. We employed an open technique by using Librosa, a Python package designed for audio analysis, to separate vocals from accompanying instrumentation. As shown in Fig. 2, there is a clear difference between the audio waveforms of the human voice and the smart speaker. The waveform of the latter is more regular, with a frequency of no less than 100 Hz. Subsequently, we deleted the segment containing the smart speaker's voice.

3) Heart Rate: Unlike heartbeat extraction methods commonly used in related studies, such as skin-approach sensors or smartwatches, we used a non-contact heart rate measurement method, remote heart rate measurement (rPPG) [12]. Remote heart rate assessment is more convenient for participants than wearing sensors for extended periods. For specific details, we refer to the Eulerian video magnification (EVM) technique [13] for heart rate estimation, which is very effective for non-contact, non-interference, and non-invasive monitoring. EVM is typically applied to RGB video to amplify small changes in skin color due to changes in blood flow to estimate heart rate [14].

TABLE I: Heart Rate Features

Feature	***Dim***	***Algorithm or Explanation***
Heart rate sequence	9	*heart rate sequence*
Max & min	2	*max,min(Heart rate sequence)*
Heart rate range	1	*max - min*
Mean	1	$\sum_{i=1}^{n} hr_i \ /n$
Mean absolute deviation	1	$\sum_{i=1}^{n} \lvert hr_i - Mean \rvert \ /n$
Root mean square	1	$\sqrt{\sum_{i=1}^{n} hr_i^2 \ /n}$
Standard deviation	1	$\sqrt{\sum_{i=1}^{n} (hr_i - Mean)^2 \ /n}$
Coefficient of variation	1	*Standard deviation / Mean*
Relative increase$_i$	8	$hr_{i+1} - hr_i$
Relative change$_i$	8	$hr_{i+1} \ / \ hr_i$
Relative increase rate$_i$	8	$(hr_{i+1} - hr_i) \ / \ hr_i$
Directional change index$_i$	8	$1 \ if \ hr_{i+1} > hr_i \ else \ 0$

* n means the length of the heart rate sequence.
* hr_i means heart rate in each window.

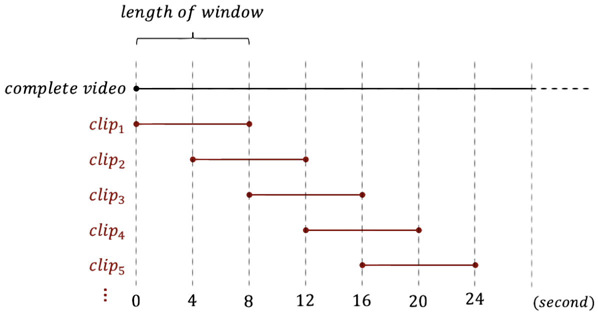

Fig. 3: Sliding window

Considering that the rPPG tool we used can detect the average heart rate throughout the video, we used the sliding window method to detect the temporal dynamics of the center rate signal of the 40-second video. As shown in Fig. 3, dividing the video into multiple overlapping 8-second segments. With this approach, we can analyze the temporal patterns and fluctuations of the signal throughout the duration of the video. After segmenting the video, we calculated different statistical features such as standard deviation, root mean square, and relative rate of increase from the heart rate values obtained during these different time intervals. Table 1 shows a total of 49 different features, and Table I summarizes all the feature sets.

C. Machine Learning Model

After extracting all the necessary features, we use two regression models for training and comparison, taking into account the presence and high dimensionality of multimodal

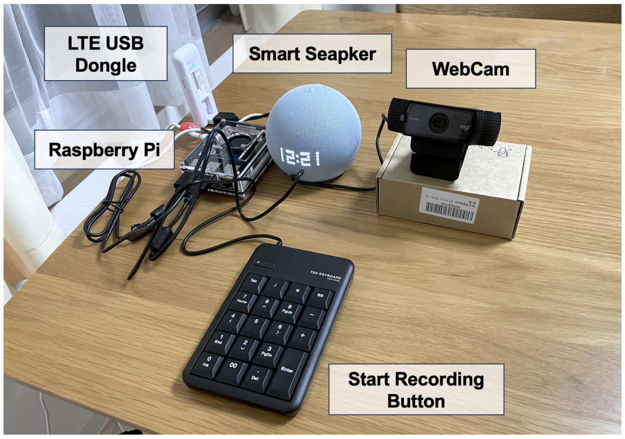

Fig. 4: Devices

features. These models include Random Forest Regression (RFR) and Light Gradient Booster (LightGBM).

RFR typically exhibits good robustness to high-dimensional data, can handle a large number of features, is not prone to overfitting, and has good modeling capabilities for nonlinear features. In addition, RFR can handle feature selection problems, is insensitive to outliers, and usually shows good generalization ability without extensive data preprocessing. LightGBM is usually good at handling high-dimensional data and nonlinear problems. It is a gradient-boosting tree method and is insensitive to outliers.

Following this, depending on the model, we conduct cross-validation with hyperparameter tuning in some of them to find out the optimal result of each model in predicting the levels of three different distress emotions.

IV. Experiment

A total of seven Japanese participated in the experiment. The experiment was conducted in the participants' respective dormitories or homes. To ensure that the brightness of the videos was not so dark that facial expressions and heart rate features could not be detected, we asked participants to maintain at least a certain level of brightness in the room during the recording.

A. Device

We used a Raspberry Pi 4 microcomputer to install the system. For the recording work, we used FFmpeg, a powerful audio-video file recording tool available on Linux. We conducted preliminary tests to verify the synchronization between audio and video in the MP4 file. By using claps and flashes, we determined that the audio was approximately 0.45 seconds ahead of the video. Finally, we used itsoffset option in FFmpeg to align them. After each recording, we utilize the PyDrive library to automatically upload files to the cloud server. A Long-Term Evolution (LTE) Dongle device is for providing internet connectivity for file uploads. In terms of other devices, we chose the Amazon Echo Dot smart speaker, C920n PRO HD webcam, and installed a button as the start button for recording, shown in Fig. 4.

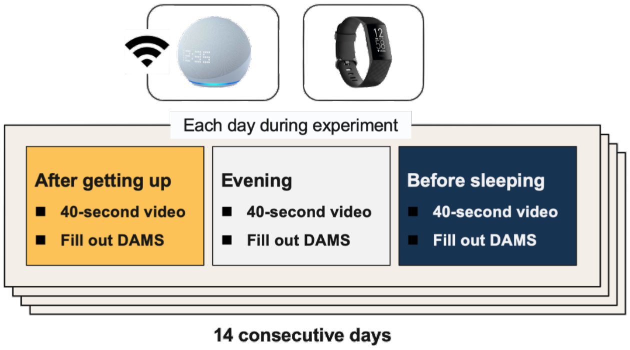

Fig. 5: Schedule

B. Questionnaire

We chose Depression and Anxiety Mood Scale (DAMS) [15] as the questionnaire in the experiment, which is specifically designed to assess the levels of happiness, depression, and anxiety distress. There are three categories of emotional words, each with three items, and each option ranges from level 0 to level 6. Besides, we also asked the subjects to wear a Fitbit watch to record physiological signals during the experiment, the purpose of which was to check the accuracy of heartbeat counts of our rPPG model by taking the value of the Fitbit watch as the true value.

C. Schedule

Participants are actively engaged in conversations with the smart speaker, ensuring a minimum of two interactions each day. These interactions occur both in the morning, shortly after waking up, and in the evening, just before bedtime. The communication typically lasts 40 seconds, encompassing two rounds of question-and-answer interactions. There are no constraints imposed on the content or topics of these interactions, making participants feel closer to their daily lives. The process spans a duration of two weeks, and at the conclusion of each conversation, participants are kindly requested to complete the Depression and Anxiety Mood Scale (DAMS) questionnaire.

V. Result

In this section, we compare the accuracy of our rPPG model against the true values obtained from heart rate data recorded by the Fitbit. Then delve into the comparison of Random Forest Regression model and LightGBM Regression model in predicting happiness, depression, and anxiety levels. Model evaluation methods encompass 10-fold cross-validation, Leave-One-Person-Out cross-validation, and Leave-One-Day-Out cross-validation for each participant.

A. Accuracy of Heart Rate

In the proposed methodology mentioned above, we utilize the EVM technique to implement rPPG to predict a participant's heart rate using the video data. To validate the accuracy of this approach, we have chosen the heart rate data

TABLE II: MAE / RMSE of Random Forest model

	Happiness	Depression	Anxiety
10-Fold CV	2.30 / 2.83	2.85 / 3.40	2.60 / 3.15
LOPO CV	2.34 / 2.87	2.71 / 3.36	2.98 / 3.67
LODO CV	2.10 / 2.69	2.59 / 3.20	2.43 / 2.97

TABLE III: MAE / RMSE of LightGBM model

	Happiness	Depression	Anxiety
10-Fold CV	2.26 / 2.84	2.97 / 3.55	2.70 / 3.34
LOPO CV	2.61 / 3.24	2.85 / 3.50	2.94 / 3.61
LODO CV	2.04 / 2.63	2.64 / 3.28	2.31 / 2.91

detected by the Fitbit AltaHR device worn by the participants. This device periodically records heart rate data at intervals of approximately 15 seconds. We utilized the Fitbit API to match the recording time of the video, access the heart rate values within that time frame, and calculate their average. Regarding the assessment of accuracy, the MAE and RMSE were 7.77 and 9.83.

B. Prediction Results for Three Distress Levels

In this subsection, we analyze the predicted results for three distress levels using the Random Forest Regression model and the LightGBM Regression model. We also present the results from three different cross-validation approaches.

Regarding the prediction results for Happiness, Depression, and Anxiety, Table II and Table III display the (MAE / RMSE) values from 10 loops of 10-Fold Cross Validation, Leave-One-Person-Out Cross Validation and Leave-One-Day-Out Cross Validation for both the Random Forest Regression model and the LightGBM Regression model, respectively.

VI. Discussion

Based on Table II and Table III, we can infer that the accuracy of distress levels is ranked from highest to lowest as follows: Happiness, Anxiety, and Depression. Additionally, for each emotion level, the lowest MAE and RMSE are consistently observed during leave-one-day-out cross-validation. LightGBM performs better than Random Forest in terms of both happiness and anxiety. However, accuracy still needs to be enhanced. Here we first explore why LODO CV outperforms the other cross-validation methods and compare the performance of the two regression models. Subsequently, we discuss the factors contributing to the overall model accuracy.

A. Subjective Differences in Participants' Perceptions

The DAMS questionnaire employs simple adjective-based questions, leading to greater variations in individual responses. Fig. 6 and Fig. 7 depict distress level variations for two of the participants in the LightGBM model's LODO CV, considering temporal changes.

Cross-validation methods such as 10-fold and LOPO CV utilize data from multiple subjects in the training phase, making it challenging to establish distinct scoring criteria for each individual. However, LODO CV is trained on a single subject's data, often achieves higher accuracy. Observing Fig. 6 and Fig. 7 above, we notice that despite the temporal perspective, they can effectively capture the general trends and fluctuations in distress levels.

B. Ablation

To understand the importance of different modalities, we conducted an ablation analysis, where we calculated MAE and RMSE by varying the combinations of facial expression, voice, and heart rate. The results are shown in Table IV; we found that for Happiness and Depression, the importance of different modalities decreases in the order of Face expression, voice, and heart rate, and for Anxiety, it is Face expression, heart rate, and voice. However, despite training the model using only the most important modality, combining all three modalities still results in the lowest MAE and RMSE.

C. Comparison of Regression Models

In LODO CV, LightGBM Regression Model demonstrated a higher incidence of lower MAE and RMSE than Random Forest Regression Model. This superiority stems from some key elements within the LightGBM framework. Firstly, LightGBM presents a lower risk of overfitting due to its leaf-wise tree growth strategy, which enables it to construct deeper and more intricate trees efficiently. In contrast, Random Forest's level-wise tree growth approach may limit its ability to capture complex patterns. Secondly, LightGBM incorporates automatic feature selection, effectively identifying and prioritizing influential features in high-dimensional data, thus enhancing model accuracy. Conversely, Random Forest often requires post-hoc analysis to determine feature importance. Furthermore, LightGBM offers superior scalability, particularly with large datasets. Its efficient handling of high-dimensional data allows it to process and analyze complex datasets with ease, further contributing to its superior performance in cross-validation experiments.

D. The Accuracy in Heart Rate Prediction

The accuracy of heart rate prediction using EVM was not good, with an MAE of 7.77. The first reason is the lighting conditions, some participants' rooms tend to have significant sunlight exposure in the morning, despite the curtains being drawn. This results in inconsistent light intensities during morning and evening video recordings, which could potentially affect the accuracy of heart rate detection. The second reason is the limitations of 2D CNN, EVM is a common 2D CNN tool for implementing rPPG. However, traditional 2D CNN approaches lack the capacity to grasp the temporal contextual aspects of facial sequences. On the contrary, 3D CNN [16] has the capability to simultaneously analyze both the spatial and temporal attributes of videos, aligning well with the characteristics of rPPG signals. This is advantageous for remote heart rate measurement.

VII. Conclusion

In this study, we focused on enabling users to monitor changes in their distress levels and proposed a multimodal

TABLE IV: Ablation

Modality Combinations	MAE / RMSE of Happiness	MAE / RMSE of Depression	MAE / RMSE of Anxiety
Face + Voice + Heartrate	2.04 / 2.63	2.59 / 3.20	2.31 / 2.91
Face + Voice	2.14 / 2.79	2.73 / 3.27	2.53 / 3.12
Face + Heartrate	2.29 / 2.87	2.78 / 3.31	2.49 / 3.09
Voice + Heartrate	2.27 / 2.84	2.85 / 3.40	2.66 / 3.23
Face	2.33 / 2.91	2.84 / 3.40	2.62 / 3.24
Voice	2.76 / 2.95	2.88 / 3.48	2.76 / 3.39
Heartrate	2.83 / 3.01	2.95 / 3.58	2.68 / 3.27

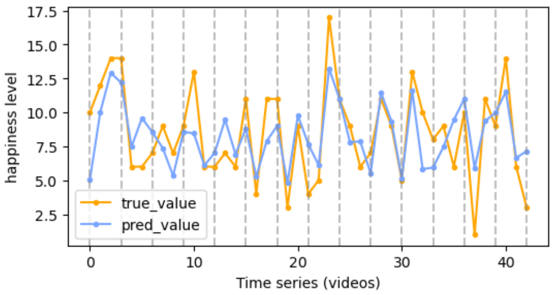

Fig. 6: Happiness Variations for Subejct 2 in LODO CV

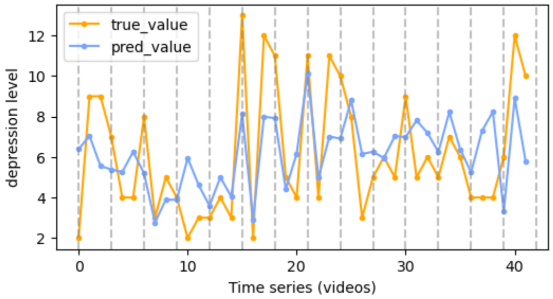

Fig. 7: Depression Variations for Subejct 1 in LODO CV

approach that utilizes imagery, audio signal, and heart rate data to estimate the levels of Happiness, Depression, and Anxiety. The data for this approach was collected from interactions between individuals and smart speakers. We evaluated the performance of different machine learning regression models and explored the reasons behind their results. The results show that MAE and RMSE were lowest with LODO CV. When performing the aggregation of scatter plots for prediction results from all subjects, Happiness and Anxiety achieved better results with LightGBM, with MAE of 2.04 and 2.31 and RMSE of 2.63 and 2.91, respectively. For future work, we plan to extend the duration of experiments with participants to gather additional data may mitigate the risk of overfitting. Furthermore, investigating alternative rPPG tools to enhance heart rate calculation precision stands as another avenue for future research.

Acknowledgments

This study was supported in part by JST PRESTO under Grant No. JPMJPR2039.

References

[1] M. A. Ozdemir, B. Elagoz, A. Alaybeyoglu, R. Sadighzadeh, and A. Akan, "Real time emotion recognition from facial expressions using cnn architecture," in *2019 Medical Technologies Congress (TIPTEKNO)*, 2019, pp. 1–4.

[2] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2016.

[3] D. D. Duncan and G. Shine, "Facial emotion recognition in real time," 2016.

[4] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 94–101.

[5] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 200–205.

[6] A. Slimi, M. Hamroun, M. Zrigui, and H. Nicolas, "Emotion recognition from speech using spectrograms and shallow neural networks," in *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia*, 2021, p. 35–39.

[7] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos)," *IEEE Access*, vol. 7, pp. 57–67, 2019.

[8] E. Smith, P. Sumner, C. Hedge, and G. Powell, "Smart speaker devices can improve speech intelligibility in adults with intellectual disability," *International Journal of Language & Communication Disorders*, vol. 56, pp. 583–593, 2021.

[9] S. Kim and A. Choudhury, "Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study," *Computers in Human Behavior*, vol. 124, p. 106914, 2021.

[10] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 59–66.

[11] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[12] X. Niu, H. Han, S. Shan, and X. Chen, "Continuous heart rate measurement from face: A robust rppg approach with distribution learning," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 642–650.

[13] Y. S. Dosso, A. Bekele, and J. R. Green, "Eulerian magnification of multi-modal rgb-d video for heart rate estimation," in *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2018, pp. 1–6.

[14] A. R. M, Sahana, V. R, S. G, and S. N, "Heart rate detection through eulerian video magnification of face videos," in *International Advanced Research Journal in Science, Engineering and Technology*, vol. 10, 2023, pp. 486–492.

[15] I. Fukui, "The depression and anxiety mood scale (dams): Scale development and validation," *Japanese Association of Behavioral and Cognitive Therapies*, vol. 23, no. 2, pp. 83–93, 1997.

[16] M. Hu, F. Qian, D. Guo, X. Wang, L. He, and F. Ren, "Eta-rppgnet: Effective time-domain attention network for remote heart rate measurement," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.