Detecting Distress Variations Using Multimodal Data Obtained through Interaction with A Smart Speaker

Chingyuan LIN[†], Yuki MATSUDA[†,††], Hirohiko SUWA[†,††], and Keiichi YASUMOTO[†,††]

† Nara Institute of Science and Technology　8916-5 Takayama-cho, Ikoma, 630–0101 Japan
†† RIKEN Center for Advanced Intelligence Project　1-4-1, Nihonbashi, Chuo-ku, Tokyo, 103–0027 Japan

Abstract Mental health significantly affects people, with excessive stress potentially causing depression, low productivity, and suicidal thoughts. It can also harm physical health, impacting appetite and sleep, and may lead to other diseases. In most cases, individuals do not notice stress buildup until their health severely deteriorates. Thus, daily monitoring of stress levels is essential. In this study, we aim to realize a method to estimate people's distress levels in everyday life through conversation with a smart speaker. We set up a smart speaker in the bedrooms of participants to simulate a home environment and recorded their interactions with it using a webcam. These recordings allowed us to analyze facial expressions, voice, and heart rate data. We processed these features and predicted levels of Happiness, Depression, and Anxiety. Participants completed questionnaires using the Depression and Anxiety Mood Scale (DAMS) after each session, providing emotion labels with scores from 0 to 18. In a 14-day experiment involving seven participants aged 22-24, the MAE for Happiness, Depression, and Anxiety levels were 2.04, 2.59, and 2.31, respectively, while the RMSE for these distress levels were 2.63, 3.20, and 2.91.

Key words distress, happiness, anxiety, depression, stress, audio-visual, heart rate, multimodality, smart speaker, DAMS.

1. Introduction

In recent years, the shift from office-based to remote work has become more prevalent. By 2023, 12.7% of full-time employees work from home, with 28.2% in a hybrid model. Tokyo has seen over half of its businesses adopt remote work, a significant increase from less than 30% pre-COVID-19, maintaining a rate above 50% by the end of 2022. While remote work reduces physical health risks associated with outdoor jobs, concerns about the psychological well-being of workers have escalated. Issues like psychological stress and depression are increasingly prominent due to the indoor office environment. Psychological conditions, unlike physical illnesses, develop gradually and are harder to detect early on. Despite research on various treatments [1], the subtle nature of mental health conditions makes early detection challenging, highlighting the need for early prevention and intervention.

The study focuses on monitoring daily emotional distress indicators in users to help them understand their mental state. It utilizes facial expressions, acoustic information, and heart rate to measure sadness, depression, and anxiety levels in participants. Data is collected in the participants' homes to ensure comfort and gather authentic data. Smart speakers, common in households for providing information and aiding in health and education, will be a central tool in this experiment. These devices, used briefly and frequently by various age groups, will be employed to record and analyze communication patterns with participants.

In our study, we gathered 40-second video recordings of 7 individuals interacting with a smart speaker for 14 days. We extracted facial expressions, voice, and heart rate features from these videos. Using Random Forest and LightGBM Regression Models, we predicted changes in Happiness, Depression, and Anxiety levels. The labels, ranging from 0 to 18, were based on the Depression and Anxiety Mood Scale (DAMS) questionnaires completed post-recording. We evaluated our models using 10-Fold Cross Validation, Leave-Person-Day-Out Cross Validation and Leave-One-Day-Out Cross Validation, assessing accuracy with Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The lowest MAE scores for Happiness, Depression, and Anxiety were 2.04, 2.59, and 2.31, respectively, while the lowest RMSEs were 2.63, 3.20, and 2.91.

2. Related Work

To predict the categories of emotions such as stress, anxiety, and happiness, many studies have used different methods to improve accuracy. Considering the following experiment procedure and model training, we categorize these diverse studies based on three distinct focal points for discussion. In Section 2.1, we list some of the studies focusing on the utilization of audio-visual data to predict emotions. Moving on to Section 2.2, we explore how previous studies obtained physiological signal information and how to utilize its representations. In Section 2.3, we summarize the problems in these related works and explore suitable research approaches.

2.1 Emotion Recognition Using Visual Data

Facial expression is one of the features commonly used to analyze emotion or stress levels due to their ability to convey nuanced non-verbal cues. Analyzing visual content enhances the depth and accuracy of emotional assessment, making it a valuable resource in psychological research. Mohammad 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]. Dagar *et al.* introduces needs and applications of facial expression recognition, gives a brief introduction towards techniques, applications, and challenges of automatic emotion recognition system [3]. In their framework, they have adeptly utilized frames extracted from live streaming, applying an advanced Grabor feature extraction technique. This is further coupled with a sophisticated Multi-Layer Perceptron (MLP) neural network. And Mehmet Akif Ozdemir *et al.* propose a low-cost and functionality method for real-time classification of seven different emotions by facial expression based on LeNet CNN architecture [4]. They merged 3 datasets (JAFFE, KDEF, and their custom dataset) and achieved an accuracy of 96.43% and validation accuracy of 91.81% for the classification of 7 different emotions through facial expressions by their CNN model-based LeNet architecture.

2.2 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 [5] visually represent the timbre, pitch, and rhythm of a voice, transform audio signals into visual data, revealing patterns and intricacies in spoken language that are key to identifying emotional states. This makes it an important tool for the analysis of emotional states and levels. Popova *et al.* consider and verify a straightforward approach in which the classification of a sound fragment is reduced to the problem of image recognition, and the waveform and spectrogram are used as a visual representation of the image [6]. The experiment was done based on the Radvess open dataset, including 8 different emotions, and half of which are negative emotions, then used the VGG-11 convolutional neural network as the image classifier. The result of this experiment was 71% instead of 12.5% accuracy for a random choice.

2.3 Emotion Recognition Using Physiological Signal

In addition to facial expressions and audio data, physiological signals like heart rate variability (HRV) and electrocardiogram (ECG) patterns are crucial in predicting the categories of emotions or emotional indices. HRV and ECG waveform can detect subtle emotional changes, often imperceptible to the naked eye. This objective data offers a reliable basis for emotional analysis. To analyze physiological signals obtained from various devices or sensors, Santamaria-Granados *et al.* employed a deep learning approach using a Deep Convolutional Neural Network (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. Sriramprakash *et al.* focused on utilizing heart rate and Galvanic Skin Response (GSR) information to extract features indicative of stress levels in working individuals [8]. They applied K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) algorithms to classify the extensive open dataset SWELL-KW. Their results showed accuracies of 66.52% and 72.82%, respectively.

The studies mentioned above used public datasets with different kinds of sensors to track things like the participants' heart signals or heartbeat counts. Some of these studies used sensors that people had to wear close to their bodies for a long time. It made the participants uncomfortable and less interested in being part of the experiment, and also complicated research requiring frequent daily data collection.

2.4 Research Directions

The issues presented in the above-mentioned related work mainly focus on the experiments and their data collection or environment. 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 participants' daily activities

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

3. 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.

3.1 Assumed Environment and Data Collection

The study was conducted in participants' private rooms, chosen for their ability to elicit more authentic emotional expressions at home. This also simplified the experimental setup by avoiding the need for a fixed external location. Participants were given the freedom to choose the content and timing of their conversations, unlike controlled experiments with designated videos or structured dialogues. This approach enhanced comfort and authenticity in emotional expression. Smart speakers, useful in health and education [9]〜[11],

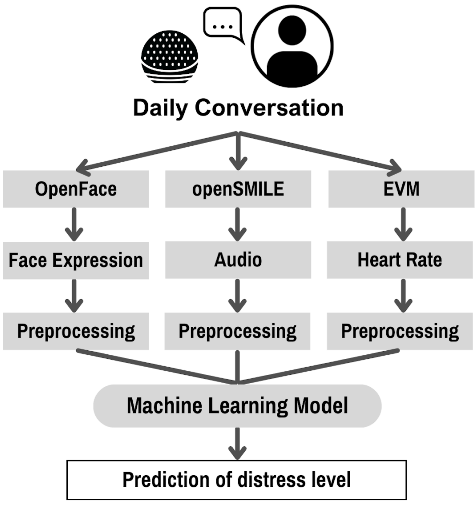

Fig. 1: Process of Data Analysis

were central to the study, providing daily information and being user-friendly across age groups [12], [13]. 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. This brief recording duration was chosen to maintain participant comfort and encourage their willingness to participate.

3.2 Feature Extraction

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

3.2.1 Face Expression Feature

For the face expression, we utilized the OpenFace toolkit, an open-source tool for facial behavior analysis presented by Tadas Baltrusaitis *et al.*, 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. For our study, we focused on extracting 6 eye-gaze and 23 facial action unit features, like Inner Brow Raiser and Blink, from 40-second video clips. Additionally, we calculated the mean and standard deviation of these 29 features.

3.2.2 Acoustic Feature

For the acoustic features in the video clips, we use the OpenSMILE open-source tool for the extraction presented by Florian Eyben *et al.*. It unites feature extraction paradigms from speech, music, and general sound events with basic video features for multi-modal processing [14]. We choose extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [15] as the feature set, which contains a total of 88 parameters. Because the webcam captures both the participant's voice and that of the smart speaker during recordings. We employed an open technique by using Librosa to separate human voice and smart speaker's voice. As shown in the spectrogram in Fig. 2, there are noticeable distinctions between the waveforms of human voices and the audio from the smart speaker. The latter exhibits a more regular waveform pattern, with a frequency that does not dip below 100 Hz. We exclude the segments of sound originating from the smart speaker, retaining only the human vocal data for training OpenSMILE.

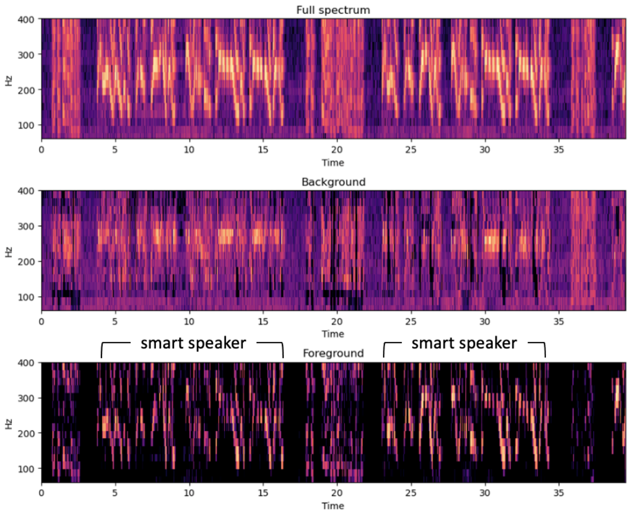

Fig. 2: Vocal Separation

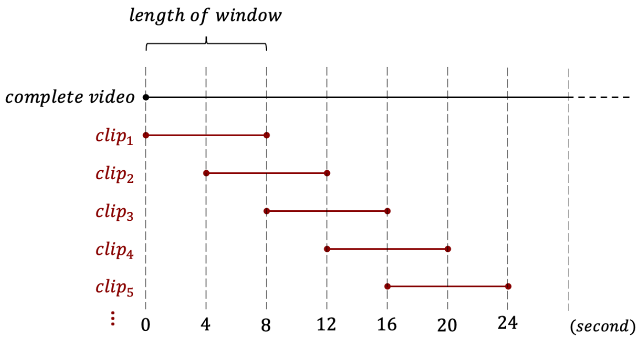

Fig. 3: Heart Rate Segmentation

3.2.3 Heart Rate

Unlike common heartbeat extraction methods using skin-close sensors or smartwatches, our study employs non-contact heart rate measurement using remote photoplethysmography (rPPG) [16], which is more convenient for participants than wearing sensors. We use the Eulerian Video Magnification (EVM) technique [17] for heart rate estimation [18]. EVM, applied to RGB video, amplifies subtle skin color changes due to blood flow variations, enabling non-contact, unobtrusive heart rate monitoring. To analyze the heart rate dynamics in a 40-second video using rPPG, we employed a sliding window method, breaking the video into overlapping 8-second segments, shown in Fig. 3. This allowed us to observe changes in the heart rate signal over time. We then extracted various statistical features, like standard deviation and root mean square, from these segments. A comprehensive list of the 49 features is in Table 1. Combining the 58 features from face expression and the 88 features from voice, we have a total of 195 feature dimensions.

3.3 Machine Learning Model

This paper discusses using two regression models, including Random Forest Regression (RFR) and Light Gradient Boosting Machine

Table 1: 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.

(LightGBM), for training and comparison in a context involving multimodal features and high dimensionality. RFR is highlighted for its robustness to high-dimensional data, resistance to overfitting, ability to handle a large number of features, often showing good generalization without much data preprocessing. LightGBM is noted for its effectiveness in high-dimensional, nonlinear problems, being a gradient-boosting tree method that's also not sensitive to outliers.

4. Experiment

There were a total of seven Japanese participants in the experiment. The experiments were conducted in the participants' respective dormitories or homes. To ensure that the brightness of the video would not be so dark that facial expression and heart rate features could not be detected, we requested the participants to maintain at least a certain level of illumination in the room during recording.

4.1 Device

We used a Raspberry Pi 4 microcomputer to install the system, and within it, and employed FFmpeg to record audio-video file. FFmpeg is suitable for various audio and video processing tasks, ranging from decoding and encoding to streaming. After each recording, we utilize the PyDrive library to automatically upload files to the cloud server. And 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 which supports 1080p high definition, and installed a button as the start button for recording. All devices as shown in Fig.4.

4.2 Questionnaire

We chose Depression and Anxiety Mood Scale (DAMS) [19] as the questionnaire in the experiment, which is specifically designed to assess the levels of happiness, depression, and anxiety distress. It is suitable for daily multiple entries and uses simple adjectives for descriptions, there are three categories of emotional words, each with three items, and each option ranges from level 0 to level 6.

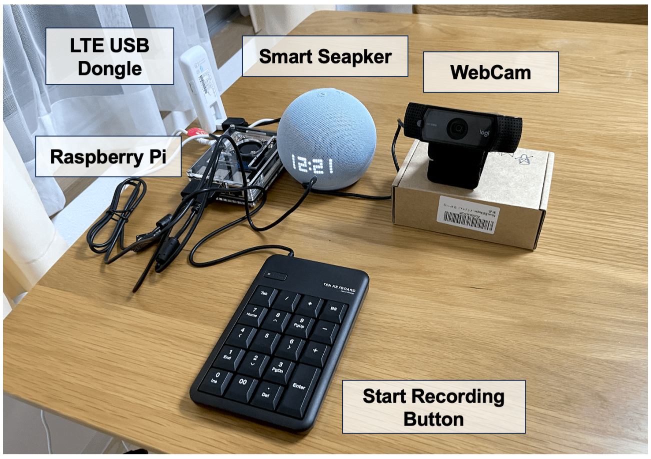

Fig. 4: Experimental Devices

Besides, we asked participants to wear a Fitbit AltaHR device to record physiological signals, the purpose of which was to verify the accuracy of our rPPG model in calculating heart rates after the experiment.

4.3 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 last 40 seconds, encompassing two rounds of question-and-answer interactions. There are no constraints imposed on the content or topics of these interactions, make 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.

5. Result

In this section, we begin by comparing the accuracy of our remote photoplethysmography (rPPG) model against the true values obtained from heart rate data recorded by the Fitbit watches worn by the participants. Then delve into the comparison of the Random Forest Regression model and the LightGBM Regression model in predicting happiness, depression, and anxiety levels. Model evaluation methods encompass 10-Fold Cross Validation, Leave-One-Person-Out Cross Validation (LOPO) and Leave-One-Day-Out Cross Validation (LODO) for each participant.

5.1 Accuracy of Heart Rate

In our method, we applied Eulerian Video Magnification (EVM) to perform remote photoplethysmography (rPPG) for estimating heart rates from video data. We compared our results with the heart rate data from Fitbit AltaHR, which records heart rates every 15 seconds. Using the Fitbit API, we synchronized the video recording times with the Fitbit data to obtain and average the heart rate values. Our heart rate predictions were made using a sliding window technique on the video to segment and average values from different parts. We

Table 2: Accuracy within different tolerance ranges

Tolerance Range	Accuracy of Heart Rate
< 5	42.79%
< 10	67.29%
< 15	86.19%

Table 3: 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 4: 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

finally got the results with an MAE of 7.77 and an RMSE of 9.73. Additionally, we also considered tolerance ranges and presented the corresponding accuracy for different tolerance ranges in Table 2.

5. 2 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 3 and Table 4 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.

6. Discussion

Based on Table 3 and Table 4, 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. And LightGBM has better performance than Random Forest for 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.

6. 1 Subjective Differences in Participants' Perceptions

The DAMS questionnaire employs simple adjective-based questions, leading to greater variations in individual responses. Fig. 5 and Fig. 6 depict distress level variations in the LightGBM model's leave-one-day-out cross-validation for some participants. Cross-validation methods such as 10-fold and leave-one-person-out involve multiple subjects in the training set, complicating the identification of unique scoring criteria for each individual. Leave-one-day-out cross-validation, trained on a single subject's data, typically yields higher accuracy. From Fig. 5 and Fig. 6 provided above, we can find that if viewed from a temporal perspective, they can still capture the overall trends and fluctuations in distress levels.

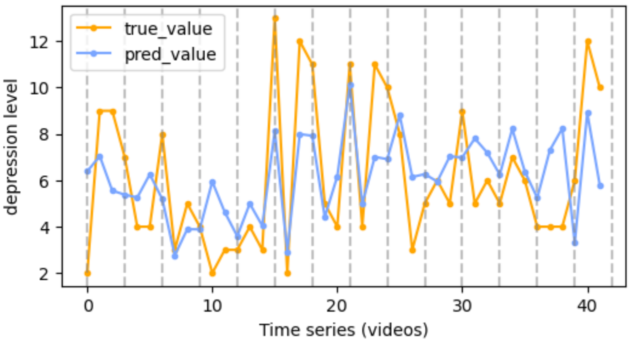

Fig. 5: Happiness Level Variations for Subejct 2 in LODO CV

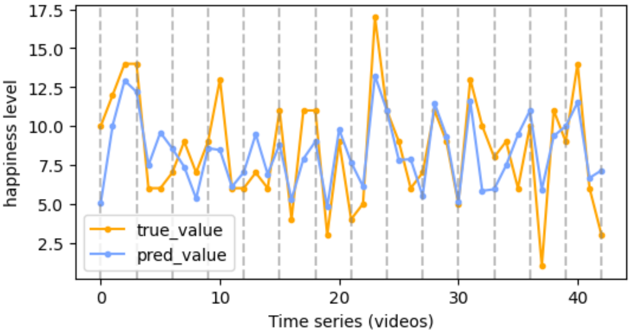

Fig. 6: Depression Level Variations for Subejct 1 in LODO CV

6. 2 Comparison of Regression Models

In leave-one-day-out cross-validation, LightGBM Regression Model exhibited a higher proportion of lower MAE and RMSE compared to the Random Forest Regression Model, there are two reasons for this result. The first one is the lower risk of overfitting, compare to Random Forest, a key advantage of LightGBM is its lower risk of overfitting, especially in datasets with many features. LightGBM's leaf-wise tree growth strategy allows it to build deeper and more complex trees efficiently, making it well-suited for high-dimensional data. Unlike Random Forest, which grows trees level-wise, LightGBM can capture intricate patterns without significantly increasing the risk of overfitting. The second one is automatic feature selection, in high-dimensional data, LightGBM efficiently identifies and prioritizes the most influential features, enhancing model accuracy in complex datasets, whereas Random Forest, while effective in many scenarios, often requires post-hoc analysis to determine feature importance.

6. 3 The Accuracy in Heart Rate Prediction

According to Table 2, the accuracy of heart rate prediction using Eulerian Video Magnification (EVM) was not good, with an MAE of 7.77. After reviewing the video data of all participants, we identified three reasons. The first one 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. Another situation is some participants tend to have backlighting during recording, resulting in insufficient lighting on one side of their faces. The second one is the distance between the webcam and participants, during the recording process, if participants move back and forth, it will lead to significant fluctuations in the predicted heart rate values. The last one is the limitations of 2D CNN, EVM is 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 [20] 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.

7. Conclusion

This study proposed a multimodal approach using imagery, audio, and heart rate data from smart speaker interactions to monitor users' Happiness, Depression, and Anxiety levels. Different machine learning regression models were evaluated, with Leave-One-Day-Out Cross Validation showing the lowest MAE and RMSE values. The LightGBM Regression Model yielded better results for Happiness and Anxiety, while the Random Forest Regression Model was more effective for Depression. In terms of future work, we should focus on longer-term experiments to gather more data and minimize overfitting risks, and explore other rPPG tools for better heart rate accuracy is also crucial. Finally, we hope this study will provide valuable insights and methods for the further development of emotion monitoring and support systems, and expect to contribute to the realization of a healthier and happier society.

Acknowledgment

This work was supported by MEXT "Innovation Platform for Society 5.0" Program Grant Number JPMXP0518071489.

Reference

[1] Z.D. Cohen and R.J. DeRubeis, "Treatment selection in depression," Annual Review of Clinical Psychology, vol.14, no.1, pp.209–236, 2018. PMID: 29494258.

[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. Dagar, A. Hudait, H.K. Tripathy, and M.N. Das, "Automatic emotion detection model from facial expression," 2016 International Conference on Advanced Communication Control and Computing Technologies, pp.77–85, ICACCCT, 2016.

[4] M.A. Ozdemir, B. Elagoz, A. Alaybeyoglu, R. Sadighzadeh, and A. Akan, "Real Time Emotion Recognition from Facial Expressions Using CNN Architecture," 2019 Medical Technologies Congress, pp.1–4, TIPTEKNO, 2019.

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

[6] A.S. Popova, A.G. Rassadin, and A.A. Ponomarenko, "Emotion Recognition in Sound," Advances in Neural Computation, Machine Learning, and Cognitive Research, eds. by B. Kryzhanovsky, W. Dunin-Barkowski, and V. Redko, pp.117–124, 2018.

[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] S. Sriramprakash, V.D. Prasanna, and O.R. Murthy, "Stress detection in working people," Procedia Computer Science, vol.115, pp.359–366, 2017. 7th International Conference on Advances in Computing and Communications, ICACC-2017, 22-24 August 2017, Cochin, India.

[9] R. Garg and S. Sengupta, "He is just like me: A study of the long-term use of smart speakers by parents and children," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol.4, pp.1–24, mar 2020.

[10] 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.

[11] D. Fedotov, Y. Matsuda, and W. Minker, "From smart to personal environment: Integrating emotion recognition into smart houses," 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom '19 Workshops), pp.943–948, 2019.

[12] 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.

[13] A. Reis, D. Paulino, H. Paredes, and J. Barroso, "Using intelligent personal assistants to strengthen the elderlies' social bonds," Universal Access in Human–Computer Interaction. Human and Technological Environments, eds. by M. Antona and C. Stephanidis, pp.593–602, 2017.

[14] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," Proceedings of the 21st ACM International Conference on Multimedia, p.835–838, 2013.

[15] 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.

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

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

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

[19] 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.

[20] 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.