Estimating Health Condition Using Facial Emotion Recognition

1st Atsunori Nishiyama
Graduate School of Engineering
Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan
atsunori@es4.eedept.kobe-u.ac.jp

2nd Takuya Nakata
Center of Mathematical and Data Science
Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan
tnakata@bear.kobe-u.ac.jp

3rd Sinan Chen
Center of Mathematical and Data Science
Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan
chensinan@gold.kobe-u.ac.jp

4th Sachio Saiki
School of Data and Innovation
Kochi University of Technology
185 Miyanokuchi, Tosayamada, Kami, Kochi, 782-8502, Japan
saiki.sachio@kochi-tech.ac.jp

5th Masahide Nakamura
Center of Mathematical and Data Science
Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan
masa-n@cmds.kobe-u.ac.jp

***Abstract*—In recent years, percentage of the elderly living alone are also on the rise. Elderly living alone are likely to be socially isolated, and if the detection of illness or injury is delayed, it is likely to lead to dying alone. It is important to understand the daily health condition of such elderly people, but existing safety confirmation and elderly care services have the problem that elderly people have no choice but to report their health condition by themselves. In this study, we use an existing emotion recognition model to acquire and analyze emotion estimation results from images of facial expressions of people acting according to their health condition. Then, we conduct an experiment to see if machine learning can classify the video clips using the emotion estimation results as feature values and the health condition of the person as labeled data. Based on the experiments, it is possible to classify facial expressions by creating a model that is adapted to the individual.**

***Index Terms*—elderly monitoring, facial expressions, facial emotion recognition, health condition, machine learning**

I. Introduction

In recent years, the proportion of elderly living alone is on the rise [1]. In such circumstances, many elderly people living alone have died at home [2], with numerous cases considered dying alone. Against this background, it is important to communicate with and watch over elderly people, and care services through visits or digital tools are being implemented [3]. However, there are issues such as a shortage of caregivers and systems where elderly people themselves have to notice and report their health condition, highlighting the need for services that check and record daily health conditions through talking. Therefore, the long-term goal of this research is to provide daily health checks through virtual agent and connect them to support, with this study focusing on investigating whether health conditions can be estimated from facial expressions.

The purpose of this study is to propose a method to estimate health conditions from facial expressions during short conversations. The key idea is to utilize existing pre-trained emotion recognition models to estimate emotions from facial expressions, develop a machine learning classification model using emotion recognition from facial expressions as features, and use short video clips simulating conversations with agent for training facial expressions. To achieve our goal, the following methods are proposed:

- A1: Collect video clips of participants acting according to their health conditions.
- A2: Extract frames from video clips and apply emotion recognition to extract emotion values for each frame.
- A3: Create features from the extracted emotion values and pair them with the correct label for each frame.
- A4: Collect the features and label of the frames to develop a machine learning classification model.
- A5: Estimate the health condition of frames in the video clip using the machine learning model and estimate the overall health condition of the clip by majority vote.

Based on this method, we actually collect video clips and estimate the health condition labels of the entire clip using a machine learning model. As a result, it was possible to estimate health condition labels by creating and using machine learning models for each participant.

II. Preliminaries

A. Increase in Elderly Living Alone

In recent years, the elderly population in Japan has been increasing, and according to FY2024 Annual Report on the Ageing Society [1], the aging rate is expected to continue rising as the population aged 65 and over increases while the total population decreases, reaching 33.3% in 2037, with one in three people being aged 65 or older. In addition, the proportion of people aged 65 and over living alone as a percentage of the population of men and women aged 65 and over is expected to increase from 15.0% for men and 22.1% for women in 2020 to 26.1% for men and 29.3% for women in 2050, indicating an increase in elderly people living alone. Regarding elderly living alone, a survey shows that the proportion of those who answered "every day" to the frequency of talking to others is less than half that of people living with others, indicating that elderly living alone are particularly isolated. In fact, in the first half of 2024, 28,330 people aged 65 and over living alone died at home [2], with numerous cases considered dying alone.

B. Key Issues to Address

To prevent the elderly from becoming isolated and their health condition from deteriorating unnoticed, it is important to talk to them daily and understand their health condition. the Guidebook for Residents to Watch Over the Elderly by the Tokyo Metropolitan Government Bureau of Social Welfare and Public Health [4] refers to the relationship where residents casually care for and watch over each other in their daily lives as "casual watching over" and states that casual watching over, where residents notice differences and abnormalities from usual and provide care or consult with specialized institutions while communicating, is becoming increasingly important.

If there is daily communication with local residents or even closer family members, they will notice differences from usual, but in isolated situations, this does not happen, so it is necessary to provide regular communication opportunities and utilize elderly care services to understand thier condition. This study focuses on daily communication and watching over.

C. Challenges in Elderly Monitoring

There are various types of elderly care services used for watching over the elderly. According to a survey by Administrative Evaluation Bureau [3], activities watching over through visits are conducted by community welfare commissioners, social welfare councils, and community comprehensive support centers to understand the living conditions through face-to-face interactions. The survey found that securing personnel for activities, such as community welfare commissioners, is becoming difficult. Visiting elderly people's homes individually requires ingenuity, and depending on the frequency of visits, it may not be possible to notice abnormalities.

Watching over using digital tools are also mentioned. When an abnormality is detected through sensors, the sensor installation company is contacted to confirm their safety [5]. Robots or tablet devices can be used to check the appearance and have conversations [6] [7]. Simply pressing the button on an emergency reporting device allows communication with the rental company [8] [3]. The challenges mentioned include local governments struggling with who will go to the house of the watched over person to confirm their safety, and some people find it cumbersome to operate touch panels to report their health condition for the day. Therefore, while digital tools can increase care time, the problem is that the elderly have to notice and report their Bad health condition themselves to convey detailed health information.

Our research group is also conducting research that watch over elderly people at home using virtual agent [9], but non-verbal information is not utilized. Given these challenges, there is a need for services that inquire about daily health conditions, utilize non-verbal information to understand health conditions, and not only understand but also record them.

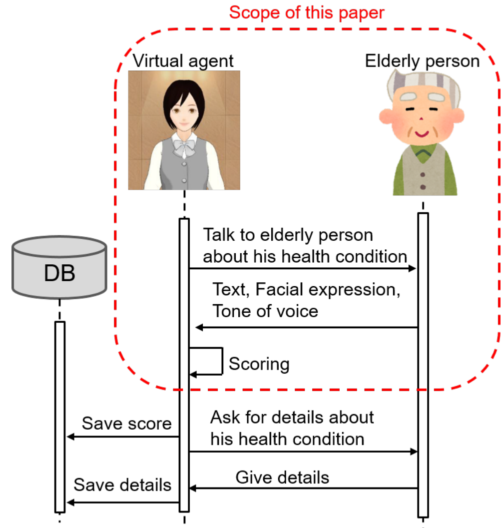

Fig. 1. Overview of Long-Term Research Goals.

D. Study Scope and Long-Term Objectives

Based on the current problem, the long-term goal of this research is to provide daily health checks through virtual agent and connect them to support. Fig. 1 shows the overall figure of the long-term goals. Virtual agent talk to an elderly person about the daily health condition, and when he or she respond, text and non-verbal information, such as facial expressions and voice tones, are obtained. The obtained information is used to analyze the health condition of the elderly and calculate it as a score. If the calculated health condition score is Bad, virtual agent will inquire about the details of the Bad condition and connect it to support. The scope of this study is to investigate whether health conditions can be estimated from facial expressions, as shown in the framed part of the long-term goals in Fig. 1.

E. Related Research

Kawamura et al. verified the effectiveness of fatigue estimation focusing on facial expression changes during speech, measured facial expressions using motion capture, and estimated fatigue using SVM with the obtained features, showing that fatigue recognition accuracy is higher based on facial expressions during speech than on standard faces [10]. Ghosh et al. conducted research on classifying pain expressions using machine learning, classified pain expressions using a system that takes the product of SVM and CNN scores, and showed that the proposed method is superior to competing methods [11]. Zhou et al. conducted research on classifying depression and apathy in elderly people with MCI using text, audio, and video, found that apathy and depression have common linguistic and facial features, features that can be used for differential diagnosis, and that the coexistence of depression and apathy shows new features [12]. Sako et al. proposed a system that analyzes emotions from the facial expressions of care recipients, measures and evaluates the effects of care as changes in the recipient's emotions [13]. Hirayama et al. proposed a method to extract facial expression changes by incorporating the concept of changing point detection to obtain scientific evidence in elderly care [14]. Chen et al. proposed a development method for an easily implementable facial recognition system by generating automatic training data through spoken dialogue agent and comparing facial features obtained using pre-trained models [15].

While research on classifying specific conditions such as pain and mental fatigue using facial expressions has been conducted, it has not delved into daily measurement and recording of health conditions for support, nor has it conducted research on inquiring about health conditions through dialogue based on the results. To conduct daily health condition understanding and recording using non-verbal information, the scope of this study is to investigate whether health conditions can be estimated from emotions inferred from facial expressions.

III. Proposed Method

A. Goal and Key Idea

The purpose of this study is to propose a method to estimate health conditions from facial expressions during short conversations. The key idea is to utilize existing pre-trained emotion recognition models to estimate emotions from facial expressions, develop a machine learning classification model using emotion recognition from facial expressions as features, and use short video clips simulating conversations with agent for training facial expressions in three conditions: Good, Normal, and Bad health. To achieve our goal, the following five approaches are proposed.

- A1: Collect video clips based on health condition.
- A2: Extract frames and apply emotion recognition to extract emotion values.
- A3: Create features from emotion values.
- A4: Develop a machine learning model.
- A5: Estimate health condition labels of the entire clip by majority vote.

The overall figure of the approach is shown in Fig. 2.

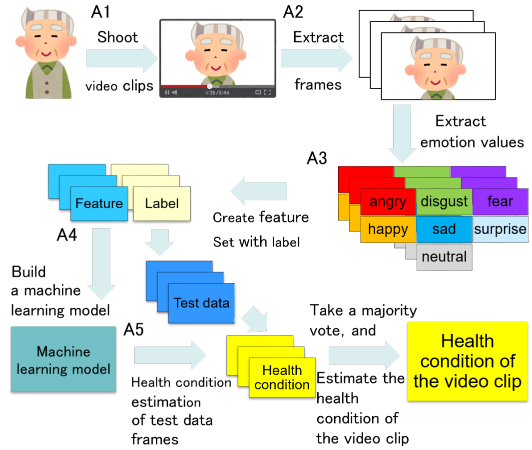

Fig. 2. Overview of Proposed Method.

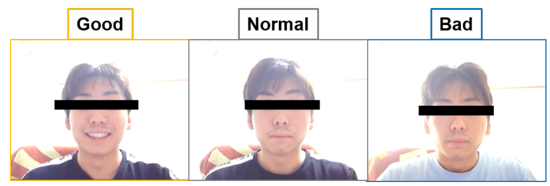

Fig. 3. Examples of Assumed Facial Expressions.

Video clip collection and emotion recognition are performed in A1 and A2, . Machine learning estimation is performed in A3, A4, and A5.

B. A1: Collecting Video Clips Based on Health Condition

First, to data for estimating health conditions, 10-second video clips of facial expressions based on health conditions are collected. When collecting video clips, three health conditions (Good, Normal, and Bad) are assumed, and one 10-second video clip is recorded for each condition. Good is assumed to be energetic and smiling, Normal is assumed to be not smiling in the extreme, but in a ceremonious state, and Bad is assumed to be the health condition when having a cold or illness or when mentally unstable. Examples of the assumed facial expressions are shown in Fig. 3, with Good, Normal, and Bad facial expressions from left to right. Based on the long-term goal, video clips simulating conversations where a virtual agent talks and participants respond are collected.

C. A2: Extract Frames and Apply Emotion Recognition to Extract Emotion Values

Frames are extracted from video clips at short intervals, and emotion values are obtained using an emotion recognition model for all frames.. This is the approach in the upeer right of Fig. 2, where many frames can be obtained from one video clip. Emotion recognition calculates the confidence of seven emotions as numerical values, referred to as emotion values.

The still images in Fig. 3 are extracted from video clips. Applying the emotion recognition model to the good condi-

TABLE I
EXAMPLE OF GOOD FRAME HAPPY PROBABILITY AND ESTIMATION RESULTS

happy	label	estimation results
0.965837	Good	Good
0.961356	Good	Good
0.961356	Good	Good
...	...	...
0.178858	Good	Normal
0.142448	Good	Normal
0.020623	Good	Bad

tion, an emotion value vector is in the order of angry, disgust, fear, happy, sad, surprise, and neutral as (0.001508, 0.000049, 0.001163, 0.961356, 0.001336, 0.000083, 0.034510). The emotion values for each frame are graphed in Fig. 4, where the solid line represents the happy values. The purpose of creating graphs is to help determine features when development a machine learning model.

D. A3: Create Features from Emotion Values

To create training data (health condition data) for estimating health conditions from facial expressions and develop a machine learning model, the emotions to be used as features are selected and paired with the correct label for the frames. Features for training the machine learning classification model are selected from the extracted emotion values. In this study, looking at Fig. 4, the happy value is particularly large in the Good condition and small in the Normal and Bad conditions, so the happy value is considered to be related to health condition, and only happy is used as a feature.

For example, if the transition of emotion values obtained from the Good video clip is the good graph in Fig. 4, the happy value of each frame in this graph is paired with good. This results in the left and middle sets in TABLE I, and this is also done for the Normal and Bad conditions, resulting in health condition data for each condition: happy value and Normal, happy value and Bad.

E. A4: Develop a Machine Learning Model

To estimate the health condition of each frame, a machine learning model is developed using health condition data. Half of the health condition data is used as training data, and a machine learning model is developed to classify frames into three health conditions. The machine learning model is developed individually for each person using only their video clips, without using other people's health condition data.

For example, to develop a machine learning model for the person in Fig. 3, health condition data like the left and middle sets in TABLE I is prepared for the emotion values in Fig. 4 obtained from the three video clips, including Normal and Bad conditions. Since 441 frames were obtained from the three video clips used in the example, using the 221-frame training data to develop a machine learning model results in a model that estimates which health condition each frame in the video clip is classified.

F. A5: Estimating Health Condition Labels of The Entire Clip by Majority Vote

To estimate health conditions from facial expressions during short conversations, the developed machine learning model is used to estimate the health condition of the test data, and the health condition of the video clip is estimated by majority vote. This is the approach in the lower of Fig. 2, where the health condition of the test data frames is estimated. By collecting the health conditions estimated from the frames of the same video clip and taking a majority vote, the health condition that appears most frequently in the video clip is considered the overall health condition of the video clip.

For example, for the person used in the previous approach example, 220 frames out of the 441 frames obtained from the three video clips are test data. Taking the Good condition as an example, there are 88 frames of test data, and estimating. As a result, Good is 63 frames, Normal is 18 frames, and Bad is 7 frames. Therefore, the most frequent label, Good, is estimated as the overall health condition of the video clip. The overall health condition of a video clip is determined by majority vote from the health conditions of its frames.

IV. EXPERIMENT

A. Experimental Setup

To collect video clips, the application program using the pre-trained model was created with reference to the sample app [16], and JavaScript was used to create an application to extract emotion values and add a function to save emotion values to a CSV file. The pre-trained emotion recognition model [17] has an accuracy of 66% on the FER-2013 emotion dataset, and the seven emotion values are calculated to sum to 1. Python was used to develop the machine learning classification model for estimating health conditions from emotion values and to evaluate the proposed method, with a decision tree model adopted for classification.

The experiment involved 12 participants, including one teenage female, nine males in their 20s, one male in his 50s, and one female in her 50s, who were researcher, his family members, and members of our laboratory. A total of 36 video clips (three per participant) were collected for the experiment. Using the three collected video clips for each participant, individual machine learning models were developed to estimate health conditions, and experiments were conducted to confirm whether the health conditions of the video clips could be correctly classified.

B. Experimental Method

First, video clips for health condition estimation were collected, and emotion values were obtained. To obtain facial expressions simulating conversations with agent, the audio “How are you feeling?” was played as talking from the agent, and the participants responded to it within 10 seconds. After 10 seconds, approximately 150 frames were extracted from the video clips, and a CSV file recording the estimated emotion values for each frame was obtained. This process of collecting video clips and obtaining CSV files was performed once for

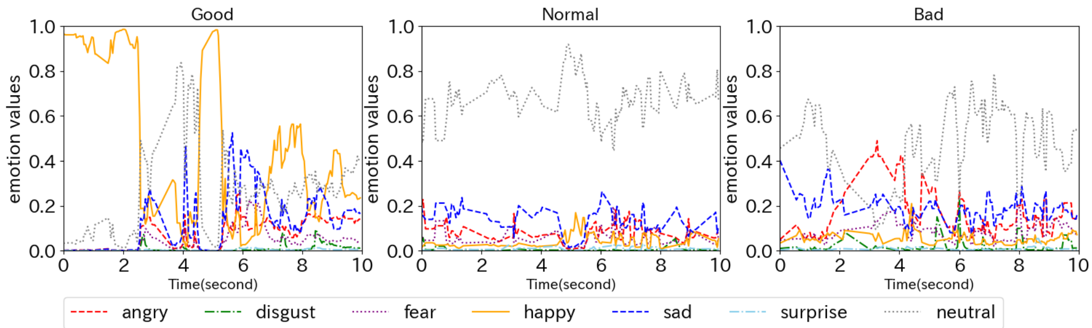

Fig. 4. Emotion Values Extracted from Assumed Facial Expressions.

each assumed health condition, and the emotion values of the facial expressions for each health conditions were obtained. The video clips and each frame used to obtain the emotion values were not saved because of privacy.

Next, a machine learning classification model was developed to estimate health conditions from the emotion values of the training data. The optimal hyperparameters for developing the model were determined using grid search, selecting the best combination of max_depth (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) and min_samples_split (2, 10, 20, 30), and the model was developed for each participant. Then, the developed model was used to estimate the health condition of each frame from the emotion values of the test data, and the most frequent label from the same video clip was determined as the estimated health condition.

To evaluate the developed model for predicting label of frames, the precision, recall, and F1-score were calculated. To evaluate the health condition estimation method for video clips by majority vote, the actual health condition and the estimated health condition were represented in a confusion matrix, and the recall, precision, and F1-score were calculated.

C. Results

The precision, recall, and F1-score for each health condition for each participant are shown in TABLE II. The confusion matrix of the majority vote results is shown in Fig. 5.

The hyperparameters for each participant, from A to L, were max_depth: 2, 2, 3, 4, 1, 1, 2, 9, 1, 5, 1, 10, and min_samples_split: 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 10. TABLE II shows high scores for Good conditions, but Normal and Bad often recorded 0, indicating misclassification.

Next, Fig. 5 show that majority voting achieved high accuracy for Good conditions (precision: 0.92, recall: 1.00, F1-score: 0.96). For the Bad condition, the precision was 0.73 and recall was 0.67, while for the Normal condition, the precision and recall were both 0.67. The F1-scores were 0.70 and 0.67.

D. Discussion

Regarding the precision and other metrics, Normal and Bad conditions were often confused with each other, resulting in extremely low values for these conditions, while Good achieved independently high values. Regarding the video clips, Fig. 5 shows that using only happy allowed more prominent smiles to be classified, resulting in particularly accurate classification for Good conditions. The machine learning classification models for each participant also showed a tendency for lower precision and other metrics for Normal and Bad conditions, leading to the result that Normal and Bad conditions were more likely to be confused with each other.

TABLE II
PRECISION, RECALL, AND F1-SCORE FOR EACH PARTICIPANT'S HEALTH CONDITION

		Precision	Recall	F-score
participant A	Bad	0.56	0.85	0.68
	Normal	0.72	0.39	0.51
	Good	0.81	0.78	0.80
participant B	Bad	0.78	0.92	0.85
	Normal	0.91	0.81	0.86
	Good	1.00	0.98	0.99
participant C	Bad	0.65	0.97	0.78
	Normal	0.86	0.43	0.58
	Good	0.93	0.95	0.94
participant D	Bad	0.78	0.37	0.50
	Normal	0.79	0.98	0.88
	Good	0.80	0.80	0.80
participant E	Bad	0.64	0.99	0.78
	Normal	0.00	0.00	0.00
	Good	0.93	0.63	0.75
participant F	Bad	0.00	0.00	0.00
	Normal	0.74	1.00	0.85
	Good	0.95	1.00	0.97
participant G	Bad	0.75	0.98	0.85
	Normal	0.98	0.64	0.78
	Good	0.99	1.00	0.99
participant H	Bad	0.60	0.62	0.61
	Normal	0.47	0.54	0.50
	Good	0.69	0.54	0.60
participant I	Bad	0.00	0.00	0.00
	Normal	0.69	0.98	0.81
	Good	0.88	0.98	0.93
participant J	Bad	0.68	0.94	0.80
	Normal	0.72	0.59	0.65
	Good	0.96	0.73	0.83
participant K	Bad	0.96	0.56	0.71
	Normal	0.00	0.00	0.00
	Good	0.57	0.97	0.72
participant L	Bad	0.54	0.43	0.48
	Normal	0.52	0.75	0.61
	Good	0.73	0.56	0.64

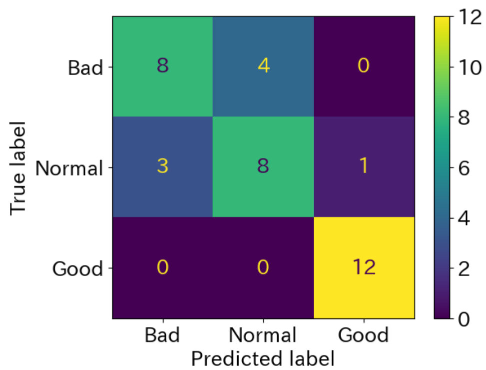

Fig. 5. Confusion Matrix of Health Condition Estimation for Video Clips.

Therefore, using only happy may not be sufficient to classify health conditions into three labels. This is because happy strongly corresponds to the Good health condition and the emotion of happiness, allowing the classification of positive reactions as Good, and using other emotion values can improve classification accuracy.

There is a possibility to improve classification accuracy, with a good balance between precision and recall, achieving an F1-score close to 0.7 and an accuracy of 0.77. The recall for Good conditions is high, and the recall for Normal and Bad conditions is about 70%, indicating that the method of creating individual models to estimate health conditions can be used. However, the video clips used in the experiment were recorded with the participants always facing the screen at a fixed angle. In real situations, the angle of the face or the camera is expected to change, so it is necessary to investigate whether this method can be used during actual conversations.

V. Conclusion

This study examined a method for estimating health conditions using emotion recognition from facial expressions. The background of this study is the increasing number of elderly people living alone, the need for daily communication, watching over, and understanding and recording health conditions, and the goal is to propose a method for estimating health conditions from facial expressions during short conversations. The approach consists of five steps: "A1: Collect video clips based on health condition", "A2: Extract frames and apply emotion recognition to extract emotion values", "A3: Create features from emotion values", "A4: Develop a machine learning model", and "A5: Estimate health condition labels of the entire clip by majority vote". Based on this proposed method, data collection, model development, and evaluation were conducted.

As a result, it was shown that individual models can be developed for each participant to estimate health conditions. Future work includes considering features that improve classification accuracy, developing a service that uses dialogue with a virtual agent to further inquire about details based on the estimated health condition, and designing a mechanism to collect data including face and camera angles and facial information other than expressions.

Acknowledgment

This research was partially supported by JSPS KAKENHI Grant Numbers JP25H01167, JP25K02946, JP24K02765, JP24K02774, JP23K17006, JP23K28091, JP23K28383.

References

[1] C. Office, "Fy2024 annual report on the ageing society," https://www8.cao.go.jp/kourei/whitepaper/w-2024/zenbun/slash06pdf_index.html, accessed April 13, 2025.
[2] N. P. A. First Investigation Division, Criminal Investigation Bureau, "Among bodies handled by police, those who died at home and lived alone - provisional figures for the first half of 2024 (jan-jun)," https://www.npa.go.jp/publications/statistics/shitai/hitorigurashi/240827_kenshi2.pdf, accessed April 13, 2025.
[3] A. E. Bureau, "Report on the results of a survey on activities to watch over elderly people living alone," https://www.soumu.go.jp/main_content/000892187.pdf, accessed April 13, 2025.
[4] T. M. G. B. of Social Welfare and P. Health, "The 4th edition of the guidebook for residents to watch over the elderly," https://www.fukushi.metro.tokyo.lg.jp/kourei/koho/jyuuminnnotameno.files/j_guidebook4.pdf, accessed April 13, 2025.
[5] I. C. Office, "Telephone monitoring service," https://www.city.isumi.lg.jp/soshikikarasagasu/fukushika/shakai_shogaifukushihan/2/1_1/1/1048.html, accessed April 13, 2025.
[6] H. Community Comprehensive Support Center Section and T. o. A. Welfare Division, "Robot for watching over people tapia price revision," https://www.town.assabu.lg.jp/uploaded/attachment/4536.pdf, accessed April 13, 2025.
[7] L. BENEFIT JAPAN Co., "Tapia," https://www.robotplanet.site/tapia/, accessed April 13, 2025.
[8] K. T. Office, "About the kumano town emergency reporting service," https://www.town.kumano.hiroshima.jp/www/contents/1422406275963/index.html, accessed April 13, 2025.
[9] H. OZONO, S. CHEN, and M. NAKAMURA, "Empirical evaluation of in-home elderly support and monitoring system using spoken dialogue agent," in *IEICE Technical Report*, vol. 123, August 2023, pp. 18–22.
[10] R. KAWAMURA, N. TAKEMURA, and K. SATO, "Mental fatigue estimation based on facial expression change during speech," in *Transactions of the Society of Instrument and Control Engineers*, vol. 53, January 2017, pp. 90–98.
[11] A. Ghosh, S. Umer, M. K. Khan, R. K. Rout, and B. C. Dhara, "Smart sentiment analysis system for pain detection using cutting edge techniques in a smart healthcare framework," in *Cluster Computing*, vol. 26, 29 January 2022, pp. 119–135.
[12] Y. Zhou, X. Yao, W. Han, Y. Wang, Z. Li, and Y. Li, "Distinguishing apathy and depression in older adults with mild cognitive impairment using text, audio, and video based on multiclass classification and shapely additice explanations," *International Journal of Geriatric Psychiatry*, vol. 37, pp. –, 09 October 2022.
[13] A. Sako, S. SAIKI, M. NAKAMURA, and K. YASUDA, "Assessing effect of care treatments using face emotion analysis of cognitive computing," in *IEICE Technical Report*, vol. 117, March 2018, pp. 105–110.
[14] K. Hirayama, S. Chen, S. Saiki, and M. Nakamura, "Toward capturing scientific evidence in elderly care: Efficient extraction of changing facial feature points," *Sensors*, vol. 21, no. 20, p. 6726, 2021.
[15] S. Chen and M. Nakamura, "Developing a facial identification system using pre-trained model and spoken dialogue agent," in *2022 International Balkan Conference on Communications and Networking (BalkanCom)*. IEEE, 2022, pp. 62–67.
[16] Pondad, "Part 17: Recognizing "emotions" with tensorflow.js," https://book.mynavi.jp/manatee/detail/id=99887, accessed April 13, 2025.
[17] "face_classification," https://github.com/oarriaga/face_classification, accessed April 13, 2025.