An IoT-Enabled, Privacy-Preserving Cyber-Physical-Social System for In-Home Caregiver Stress and Fatigue Monitoring

Sinan Chen[1,2*] and Masahide Nakamura[2,1†]

[1*]Center of Mathematical and Data Sciences, Kobe University, 1-1 Rokkodai-cho, Nada, Kobe, 657-8501, Hyogo, Japan.
[2], RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan, Country.

*Corresponding author(s). E-mail(s): chensinan@gold.kobe-u.ac.jp;
Contributing authors: masa-n@cs.kobe-u.ac.jp;
†These authors contributed equally to this work.

Abstract

The global aging trend has intensified the need for in-home caregiving, exposing caregivers to persistent physical and emotional stress. Existing monitoring solutions are largely recipient-centric and fail to address caregivers' well-being. This paper presents an IoT-enabled, privacy-preserving cyber-physical-social system (CPSS) for real-time stress and fatigue monitoring using ambient sensing. The proposed system employs commodity cameras and lightweight edge AI models to extract posture and facial landmarks, processes anonymized keypoints locally on edge devices (e.g., Raspberry Pi, Intel NUC), and applies a human-in-the-loop annotation mechanism for personalized stress classification. Unlike cloud-based or wearable-dependent approaches, the framework ensures data minimization, ethical integrity, and user autonomy while maintaining feasibility for household deployment. A modular architecture integrating sensing, edge analytics, and personalization layers demonstrates low-latency, privacy-compliant operation validated through simulated and real-world caregiving contexts. The study contributes a scalable and socially aligned CPSS design paradigm for ambient healthcare applications.

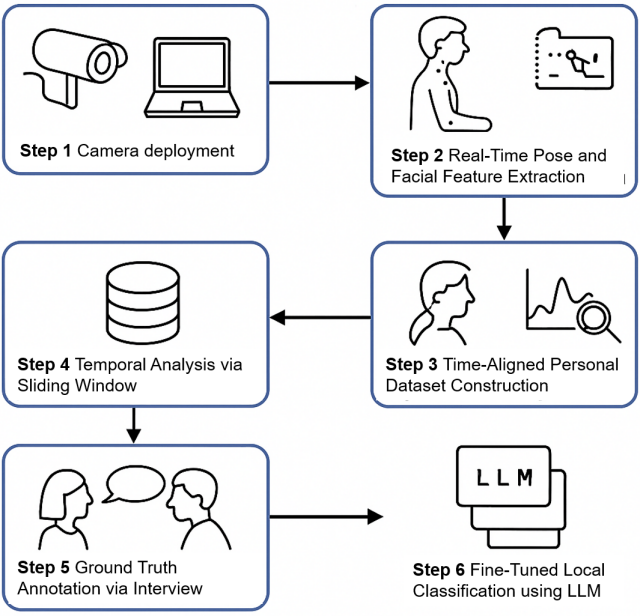

Fig. 1: Overall framework for local stress and fatigue detection.

Keywords: Internet of Things, Cyber-Physical-Social Systems, Edge AI, Privacy-Preserving Sensing, Caregiver Stress Monitoring

1 Introduction

The global aging trend has intensified demands for in-home caregiving, exposing caregivers to prolonged physical and emotional stress [1]. Unlike clinical settings, home-based care involves unique challenges such as social isolation, irregular work schedules, and strong emotional attachment to care recipients. Continuous exposure to these factors leads to cumulative fatigue, deteriorating both caregiver well-being and the quality of care [2]. Current monitoring systems, however, remain predominantly patient-centric, creating a critical gap in caregiver-oriented assessment frameworks.

Existing stress evaluation methods, such as self-reporting surveys and clinical interviews, suffer from subjectivity and recall bias [3], while wearable sensors often face low acceptance in private homes due to intrusiveness. Moreover, no existing framework effectively integrates nonverbal behavioral cues, including posture dynamics and facial micro-expressions, for real-time stress and fatigue assessment in unconstrained environments, despite their proven correlation with physiological markers [4].

To address these challenges, this study proposes a lightweight, IoT-integrated cyber-physical-social system (CPSS) for context-aware monitoring of in-home caregiver fatigue and stress. The framework captures multimodal cues: body posture and facial expressions using inexpensive, non-contact sensors embedded in everyday environments, and processes anonymized keypoints locally on edge devices. This edge–IoT architecture aligns with CPSS principles by fusing physical signals, cyber inference, and socially adaptive feedback, thereby enabling individualized, privacy-conscious monitoring. The design supports scalable deployment of ambient healthcare technologies while maintaining ethical integrity and user autonomy.

The proposed pipeline employs locally executable models such as MoveNet (body pose estimation) and MediaPipe or face-api.js (facial landmark detection) without relying on cloud infrastructure. As illustrated in Figure 1, a USB camera connected to a local device continuously captures RGB frames, which are processed in real time to extract keypoints. These are temporally aligned to build an individualized dataset, followed by a sliding-window analysis that computes statistical features to identify abnormal patterns. Post hoc interviews then assign semantic labels (e.g., "mild fatigue," "stress episode") to selected segments, after which a lightweight large language model (LLM) is fine-tuned locally for personalized stress classification. This design enables real-time, on-device analysis while preserving privacy and autonomy, representing a shift toward adaptive, human-centered monitoring for caregiving environments.

2 Preliminaries

2.1 Previous Studies

Several prior studies have explored stress mitigation strategies for elderly individuals at home, offering insights that indirectly inform caregiver-centric monitoring [5]. Horie et al. [6] proposed a video-based intervention framework that reduced stress among elderly users through personalized video playback. By collecting preference information and automatically generating recommended lists via the YouTube API, the system demonstrated the feasibility of aligning multimedia content with user interest and tracking emotional responses via facial expressions.

Later, Horie et al. [7] integrated the Rakuraku Video Service with a spoken dialogue agent (PC-Mei) [8], enabling hands-free access to personalized video content. Their two-week user study confirmed the system's effectiveness in eliciting enjoyment, relaxation, and reduced stress, emphasizing the importance of individualized, concrete user preferences.

Although these works primarily targeted elderly users, they share key motivations with the present study: (1) developing non-intrusive, personalized support systems for home environments; (2) leveraging multimodal cues (e.g., facial expressions) for implicit emotion recognition; and (3) accommodating usability constraints among non-expert users. In contrast, our research focuses

on in-home caregivers, whose stress arises from different physical and emotional sources. The proposed framework therefore shifts from content-based relaxation toward real-time behavioral sensing based on pose and microexpression dynamics, emphasizing edge computation, privacy, and personalized stress modeling.

2.2 Image-Based Human Pose Estimation

Recent advances in computer vision have produced lightweight, deployable models suitable for edge environments such as in-home caregiving scenarios [9] [10]. Among these, **PoseNet** [11] and **MoveNet** [12] achieve real-time 2D pose detection from monocular RGB images [13]. The process typically employs pre-trained MoveNet.js or PoseNet.js models that convert each frame into structured keypoints representing skeletal joints, forming the basis for posture and fatigue analysis. These models are optimized for local execution, supporting deployment on low-power devices such as Raspberry Pi or within modern browsers via WebGL and TensorFlow.js [14].

Both models take an RGB image as input and output a set of body keypoints representing anatomical landmarks with associated confidence scores. MoveNet generates 17 standardized keypoints, including the nose, eyes, shoulders, elbows, hips, knees, and ankles, which are suitable for downstream tasks such as posture classification, fatigue inference, or motion-based stress monitoring [15]. PoseNet, one of the first frameworks enabling browser-based pose estimation, offers flexible model architectures for mobile inference, while MoveNet improves speed and accuracy through TensorFlow Lite acceleration. Their public availability enables rapid integration into real-world caregiver monitoring systems.

These pose estimation models also support privacy-conscious design since all image processing can occur locally, avoiding cloud-based video transmission, an essential requirement for ethical in-home sensing.

2.3 Image-Based Human Facial Recognition

Facial recognition plays a vital role in assessing emotional states such as stress and fatigue. Two widely used frameworks for real-time facial analysis are **face-api.js** [16] and **MediaPipe Face Mesh** [17]. Both extract features such as eyebrow lift, eye openness, and mouth curvature that reflect micro-expressions linked to emotional stress. These models run efficiently on low-power platforms like Raspberry Pi and in standard browsers using TensorFlow.js or WebGL.

Given an input RGB image, these frameworks output a set of facial landmarks with 2D coordinates and confidence values. For example, **face-api.js** detects 68 landmarks across the eyes, nose, lips, and jawline, while **MediaPipe Face Mesh** produces up to 478 3D landmarks, supporting fine-grained expression analysis (e.g., smile detection or eyelid tightening) indicative of emotional fatigue [18]. The ability to run all inference locally makes them highly suitable

Table 1: Sample Output Keypoints Format

Model	Keypoints	Shape	Features
MoveNet	Body joints (e.g., nose, elbow)	17×3	(x, y, c)
face-api.js	Facial landmarks	68×3	$(x, y, confidence)$

for privacy-sensitive caregiver monitoring applications, aligning with ethical requirements for unobtrusive sensing.

Recent studies have applied these frameworks in affective computing to estimate emotional arousal and valence [19], demonstrating their scalability and robustness for long-term home deployment with minimal computational overhead.

3 Proposed Method

3.1 Flows of Designed Framework

The proposed framework monitors caregiver fatigue and stress in real time through image-based sensing that combines body posture and facial microexpression analysis. Figure 1 outlines the data flow across components. The implementation leverages lightweight, locally executable models introduced in Sections 2.2 and 2.3, ensuring that all sensing and computation occur within the home environment to protect privacy.

Step 1: Camera Deployment

A fixed-angle USB camera connected to a local edge device (e.g., Raspberry Pi or laptop) continuously captures RGB frames under ambient lighting. To reduce privacy intrusion and computational load, low-resolution (640×480) images at 5–10 FPS are used.

Step 2: Real-Time Feature Extraction

Each frame is processed in real time using MoveNet for body pose detection and MediaPipe or face-api.js for facial landmark extraction. Both output structured keypoint arrays describing skeletal and facial geometry (Table 1).

Step 3: Time-Aligned Dataset Construction

Extracted data are synchronized with timestamps for temporal fusion. Each record includes pose vectors, facial landmarks, and derived metrics (e.g., head tilt, eye aspect ratio). A sample frame format is shown in Listing 1.

Listing 1: Sample Frame Data

```
{
  "timestamp": "2025-06-01T15:43:27Z",
  "pose": [[341, 183, 0.95], [353, 240, 0.88], ...],
  "facial_landmarks": [[121, 98, 0.93], [125, 102, 0.91], ...]
```

Table 2: Annotated Dataset Example

Window	Avg. Eye Ratio	Posture Variance	Label
2025-06-01 15:40–15:45	0.19	0.03	Mild Stress
2025-06-01 15:45–15:50	0.31	0.15	Normal

```
}
```

Step 4: Temporal Analysis via Sliding Window

A moving time window (e.g., 5-minute non-overlapping segments) is applied to compute statistical features such as body sway, posture variance, or eyebrow angle change. These are aggregated for anomaly detection, as illustrated in Listing 2.

Listing 2: Feature Aggregation Code

```
window_data = get_window(dataset, start_time, end_time)
mean_eye = np.mean([eye_ratio(f) for f in window_data])
std_shoulder = np.std([shoulder_tilt(p) for p in window_data])
```

Step 5: Ground Truth Annotation

Flagged time segments (e.g., inactivity or facial strain) are reviewed in post hoc interviews. Caregivers assign semantic labels such as "normal," "mildly stressed," or "physically tired," which are stored as ground truth (Table 2).

Step 6: Local Classification using LLM

The labeled dataset is used to fine-tune a lightweight transformer-based classifier or local large language model (LLM). Each feature vector is encoded as short textual input (e.g., "unstable head tilt, raised eyebrows") and mapped to a stress category. A simplified pseudocode is shown in Listing 3.

Listing 3: Model Training Pseudocode

```
from transformers import AutoTokenizer,
    AutoModelForSequenceClassification
tokenizer = AutoTokenizer.from_pretrained("sentence-
    transformer-local")
model = AutoModelForSequenceClassification.from_pretrained("
    local-stress-detector")
train(model, tokenizer, annotated_dataset)
```

This closed-loop, privacy-preserving framework adapts to each caregiver's behavioral pattern while remaining fully local and modular for future multimodal integration.

3.2 Design Considerations

The framework provides a novel, privacy-aware approach for real-time stress and fatigue monitoring. Its **local deployability** allows all sensing and inference on low-power devices (e.g., Raspberry Pi, laptops) without cloud dependence, ensuring **data privacy** and **low-latency** feedback. The system is also **modular and extensible**: each component from acquisition to classification can be independently improved, and the human-in-the-loop annotation step supports personalized calibration.

Despite these strengths, several limitations remain. First, pose and facial data alone may not fully capture complex physiological or emotional states, especially without additional modalities such as speech or biosignals. Second, accuracy may be affected by environmental factors (lighting, occlusion, camera angle), introducing noise and missing data. Third, manual annotation is labor-intensive and may introduce bias, while local fine-tuning of LLMs, though feasible, imposes computational constraints on edge devices. Future versions could address these issues through multi-modal fusion and adaptive model compression.

From a communication perspective, the proposed CPSS operates as a closed local-area network (LAN) system, where camera nodes transmit compressed keypoint packets ($< 10\,\text{KB}$ per frame) to the processing unit via TCP/IP. This local connectivity design minimizes bandwidth usage (approximately $0.8\,\text{Mbps}$) and eliminates cloud dependency, aligning with the principles of efficient and secure embedded communication in IoT-based healthcare systems.

4 Experimental Implementation

4.1 Implementation Purpose

This implementation validates the practicality of the proposed framework and its deployment in realistic caregiving contexts. The objectives are threefold:

1. **Technical realization**: demonstrate feasibility of executing the pipeline on edge devices, highlighting real-time capability and resource efficiency.
2. **System integration**: show how stress-related features are extracted through temporal aggregation and personalized via lightweight LLM fine-tuning.
3. **Privacy-aware deployment**: verify that local processing safeguards data privacy while supporting functional caregiving applications.

These objectives emphasize implementation feasibility rather than quantitative benchmarking.

4.2 Implementation Setup

4.2.1 Hardware and Environment

The system was deployed on two representative edge platforms:

- **Raspberry Pi 4B** (Broadcom BCM2711, 4 GB RAM) for resource-constrained use.
- **Intel NUC 11** (i5-1135G7, 16 GB RAM) for mid-tier home environments.

Both ran Ubuntu 22.04 LTS with TensorFlow Lite 2.12.0. Webcam input was processed at 640×480 resolution (10 FPS) through the OpenCV pipeline described in Section 3.1.

4.2.2 Demonstration Scenarios

Two representative scenarios were implemented:

- **Simulated caregiving session**: a participant (female, age 63) performed standardized tasks in controlled conditions to verify pipeline functionality.
- **Real home deployment**: an experienced caregiver (5.2 years) used the system during daily routines, with only anonymized keypoints stored locally.

4.2.3 Annotation Protocol

Post-session interviews followed the labeling protocol in Section 3.1, Step 5:

$$\mathcal{L} = \{\texttt{normal}, \texttt{mild\_stress}, \texttt{physical\_fatigue}, \texttt{emotional\_strain}\} \tag{1}$$

This process demonstrated how multimodal streams can be mapped to interpretable categories without requiring large-scale datasets.

4.2.4 Baseline References

For contextualization, two reference approaches were included:

1. **Local-based pipeline**: MoveNet.js with face-api.js.
2. **Wearable-based**: SVM classifier using Empatica E4 biosignals.

These baselines served as qualitative references, not for performance comparison.

4.3 Implementation Highlights

The implementation confirms that:

- The full framework runs on both constrained and mid-tier edge devices without cloud dependence.
- Real-time pose and facial feature extraction is feasible under caregiving conditions.
- Privacy is preserved through exclusive local processing and anonymized storage.

Table 3 summarizes the demonstrated capabilities.

Table 3: Implementation aspects and demonstrated capabilities

Aspect	Demonstrated Capability
Technical realization	Deployment on Raspberry Pi and Intel NUC
System integration	End-to-end pipeline with temporal features and LLM-based personalization
Privacy-aware deployment	Local-only processing with anonymized keypoint storage

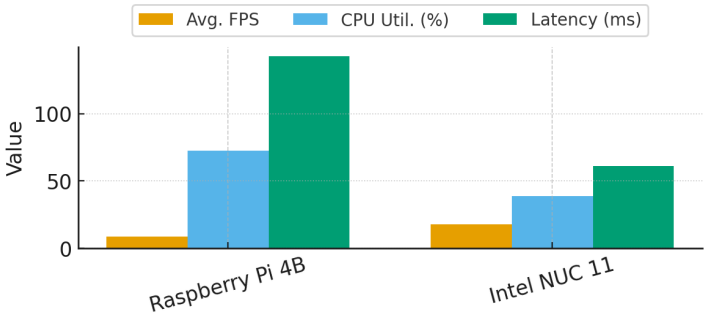

Fig. 2: Performance comparison across edge devices (higher FPS is better; lower CPU/latency is better).

Table 4: Performance comparison of edge devices

Platform	Avg. FPS	CPU Util. (%)	Latency (ms)
Raspberry Pi 4B	8.7	72.4	142.5
Intel NUC 11	17.9	38.6	61.3

4.4 Performance Evaluation

To further validate system efficiency, real-time performance was measured on both edge devices under identical settings (640×480, 10 FPS). Table 4 summarizes average frame rates, CPU utilization, and end-to-end latency, confirming the feasibility of on-device operation. Compared with conventional cloud-based stress monitoring pipelines, the proposed edge configuration reduces average processing latency by over 70 % while fully eliminating outbound data transfer. As shown in Figure 2, the Intel NUC 11 achieves nearly double the frame rate of the Raspberry Pi 4B while maintaining lower CPU utilization and latency. Figure 3 illustrates the inverse relationship between FPS and latency, confirming the advantage of higher-performance edge hardware. These results confirm that both configurations achieve real-time inference suitable for continuous caregiver monitoring while maintaining acceptable computational loads and energy efficiency.

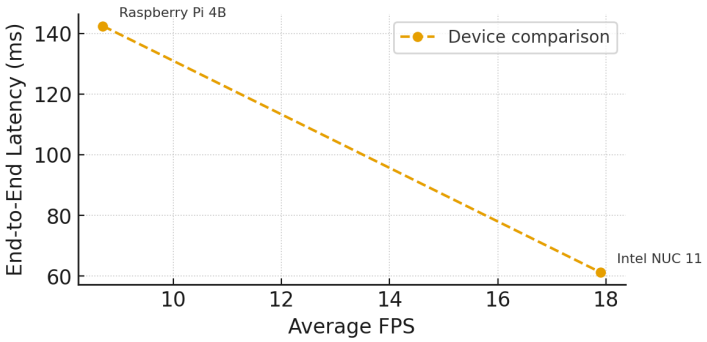

Fig. 3: Latency versus frame rate under identical settings (640×480, 10 FPS capture).

4.5 Discussion

The implementation confirms the framework's **technical feasibility** on commercially available edge devices, demonstrating low-latency multimodal processing without cloud support. The integration of pose estimation, facial analysis, and local LLM-based classification highlights the potential for adaptive, privacy-preserving stress recognition in home environments.

However, several limitations remain. The study scale is small, restricting generalizability, and results were obtained under controlled conditions that may not fully reflect real-world variability in lighting or activity. Manual annotation, while essential for personalization, introduces subjectivity, and local fine-tuning of LLMs imposes computational constraints. Future work will expand validation with larger datasets, explore hybrid edge–cloud architectures, and optimize model compression for sustained deployment.

Overall, the findings verify the framework's operational practicality and privacy compliance, establishing a foundation for scalable, stress-aware caregiving support systems. In addition to latency, bandwidth and power consumption were profiled. The average data rate per camera node was approximately 0.8 Mbps, and total system power consumption on the Raspberry Pi remained below 6.5 W during continuous inference, confirming suitability for 24/7 operation in home environments.

5 Conclusion

This study presented a locally deployable, privacy-aware framework for detecting stress and fatigue in in-home caregivers through image-based sensing. By integrating lightweight models such as MoveNet and MediaPipe for pose and facial landmark extraction, the system enables real-time, low-latency monitoring without cloud dependence, aligning with ethical requirements for data privacy in domestic environments. The main contributions are threefold: (1) demonstrating the feasibility of integrating pose and facial data into a unified

temporal dataset tailored to individual caregivers; (2) introducing a semi-supervised annotation process based on real user feedback to link objective sensing with subjective perception; and (3) implementing a modular pipeline from acquisition to personalized classification on affordable edge devices such as the Raspberry Pi.

Nevertheless, several challenges remain. The exclusive reliance on visual cues limits sensitivity to internal states such as mental fatigue or emotional burnout. Environmental factors, camera occlusion, and the computational cost of local LLM training also affect long-term scalability and robustness. Future work will focus on multimodal fusion with speech and wearable data, automated annotation through anomaly detection, and real-time intervention via dialogue agents.

Acknowledgements

This research was partially supported by JSPS KAKENHI Grant Numbers JP25H01167, JP25K02946, JP24K02765, JP24K02774, JP23K17006, JP23K28091, JP23K28383, and JST SICORP Grant Number JPMJKB2312.

References

[1] Iaboni, A., Spasojevic, S., Newman, K., Schindel Martin, L., Wang, A., Ye, B., Khan, S.S.: Wearable multimodal sensors for the detection of behavioral and psychological symptoms of dementia using personalized machine learning models. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **14**(1), 12305 (2022). https://doi.org/10.1002/dad2.12305

[2] Settanni, M., Suma, K., Adamson, L.B., McConachie, H., Servili, C., Salomone, E.: Treatment mechanism of the who caregiver skills training intervention for autism delivered in community settings. Autism Research **17**(1), 182–194 (2024). https://doi.org/10.1002/aur.3058

[3] Naegelin, M., Weibel, R.P., Kerr, J.I., Schinazi, V.R., La Marca, R., von Wangenheim, F., Ferrario, A.: An interpretable machine learning approach to multimodal stress detection in a simulated office environment. Journal of Biomedical Informatics **139**, 104299 (2023). https://doi.org/10.1016/j.jbi.2023.104299

[4] Ren, S.: Computer vision for facial analysis using human-computer interaction models. Journal of Interconnection Networks **22**(Supp03), 2144005 (2022). https://doi.org/10.1142/S0219265921440059

[5] Chen, S., Nakamura, M., Yasuda, K.: A study for estimating caregiving contexts based on extracting nonverbal information from elderly people at home. In: International Conference on Human-Computer

Interaction, pp. 259–268. Springer, ??? (2023). https://doi.org/10.1007/978-3-031-34917-1_19. https://doi.org/10.1007/978-3-031-34917-1_19

[6] Horie, H., Chen, S., Nakamura, M., Yasuda, K.: Study of stress relief service by watching personalized videos for elderly people at home. In: Proceedings of the 2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), pp. 130–136 (2022). https://doi.org/10.1109/IAICT55358.2022.9887490. https://doi.org/10.1109/IAICT55358.2022.9887490

[7] Horie, H., Chen, S., Nakamura, M., Yasuda, K.: Developing elderly stress-relief service using personalized videos and spoken dialogue agent. In: Computer Science & Information Technology (CS & IT) – CSCP 2023, pp. 1–11 (2023). https://aircconline.com/csit/papers/vol13/csit132401.pdf

[8] Chen, S., Nakamura, M.: Study of multi-modal diary service using spoken dialogue agent for self-care in elderly people. In: 2022 1st International Conference on Software Engineering and Information Technology (ICoSEIT), pp. 120–125 (2022). IEEE

[9] Chen, S., Saiki, S., Nakamura, M.: Nonintrusive fine-grained home care monitoring: Characterizing quality of in-home postural changes using bone-based human sensing. Sensors **20**(20), 5894 (2020)

[10] Chen, S., Nakamura, M.: Pose-driven dialogue framework integrating agents and locations for elderly. In: 2024 IEEE 6th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), pp. 457–459 (2024). https://doi.org/10.1109/ECBIOS61468.2024.10885449. https://doi.org/10.1109/ECBIOS61468.2024.10885449

[11] Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K.: Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 269–286 (2018). https://doi.org/10.48550/arXiv.1803.08225. https://doi.org/10.48550/arXiv.1803.08225

[12] Washabaugh, E.P., Shanmugam, T.A., Ranganathan, R., Krishnan, C.: Comparing the accuracy of open-source pose estimation methods for measuring gait kinematics. Gait & Posture **97**, 188–195 (2022). https://doi.org/10.1016/j.gaitpost.2022.06.015

[13] Chen, S., Nakamura, M., Yasuda, K.: Detecting irregular contexts using image recognition and data analysis for elderly monitoring. In: 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB), pp. 376–378 (2024). https://doi.org/10.1109/ICEIB61477.2024.10602594. https://doi.org/10.1109/

ICEIB61477.2024.10602594

[14] Chen, S., Nakamura, M., Yasuda, K.: Proposing human-centered monitoring framework characterizing contexts with vision-based edge ai. In: 2024 IEEE 7th Eurasian Conference on Educational Innovation (ECEI), pp. 309–312 (2024). https://doi.org/10.1109/ECEI60433.2024.10510862. https://doi.org/10.1109/ECEI60433.2024.10510862

[15] Chen, S., Nakamura, M.: Evaluating feasibility of pose detection with image rotation for monitoring elderly people at home. Engineering Proceedings **89**(1), 28 (2025). https://doi.org/10.3390/engproc2025089028

[16] Justad, V.: face-api.js: JavaScript API for face detection and face recognition in the browser with TensorFlow.js. GitHub Repository (2018). https://github.com/justadudewhohacks/face-api.js

[17] Lugaresi, C., Tang, J., Nash, H., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2020)

[18] Chen, S., Ozono, H., Nakamura, M., Yasuda, K.: Quantitative expression of elderly multi-modal emotions with spoken dialogue agent and edge ai. In: 2023 IEEE 6th Eurasian Conference on Educational Innovation (ECEI), pp. 219–221 (2023). IEEE

[19] Khanzada, A., Bai, C., Celepcikay, F.T.: Facial expression recognition with deep learning. arXiv preprint arXiv:2004.11823 (2020)