Thermal Camera-Based Cognitive Load Estimation: A Non-Invasive Approach

Zoja Anžur
zoja.anzur@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Gašper Slapničar
gasper.slapnicar@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Mitja Luštrek
mitja.lustrek@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Abstract

Cognitive load (CL) monitoring is a growing area of interest across various domains. Most traditional methods rely on either subjective assessments or intrusive sensors, limiting their practical applicability. In this study, we present a non-invasive approach for estimating CL using thermal imaging. Thermal videos were collected from 18 participants performing a battery of tasks designed to induce varying levels of CL. Using a low-cost thermal camera, we extracted features from facial regions of interest and trained several machine learning models, including Random Forest, Extreme Gradient Boosting, Stochastic Gradient Descent (SGD), k-Nearest Neighbors, and Light Gradient Boosting Machine, on a binary classification task distinguishing between rest and high CL conditions. The models were evaluated using Leave-One-Subject-Out cross-validation. Our results show that all models outperform the baseline majority classifier, with SGD achieving the highest accuracy (0.64 ± 0.16), despite variability across individuals. These findings support the feasibility of thermal imaging as an unobtrusive tool for CL estimation in real-world applications.

Keywords

cognitive load estimation, thermal imaging, physiological computing, machine learning for affective computing, non-invasive user monitoring

1 Introduction

Monitoring cognitive load (CL) unobtrusively and accurately has become an increasingly important goal across various domains. Traditional methods such as the NASA-TLX questionnaire [7] for assessing cognitive states often rely on intrusive sensors or subjective self-reporting, limiting their practicality in real-world applications. In recent years, the use of machine learning techniques combined with physiological signals has opened new possibilities for non-invasive and continuous monitoring [2].

The primary objective of our study was to predict CL using data obtained with a thermal camera. Our aim was to develop a method for unobtrusive measurement of physiological signals that achieves high accuracy. Compared to other physiological measurement tools, thermal cameras are relatively low-cost and quick to deploy, which makes them a practical choice for real-world cognitive monitoring applications.

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.skui.3714

2 Related Work

Early approaches of contact-free thermal monitoring of psychophysiological states based on infrared thermal imaging focused primarily on emotional and affective research [8]. Physiological background was heavily explored, specifically how autonomic nervous system activity yields descriptive thermal signatures related to affect in facial regions. Such work laid the critical groundwork for later expansion towards CL estimation.

One of the fundamental studies towards thermal-camera-based CL estimation was published by Abdelrahman et al. in 2017. They introduced an unobtrusive method that uses a commercial thermal camera to monitor temperature changes on the forehead and nose, which were chosen as regions of interest based on physiological background established earlier. It demonstrated that the difference between forehead and nose temperature correlates robustly with task difficulty, showing effectiveness in Stroop test and reading complexity experiments. Notably, the system achieved near-real-time detection with an average latency of 0.7 seconds, making it suitable for responsive, real-time cognition-aware applications [1].

While such monitoring traditionally required relatively expensive hardware [6], recent work showed potential of more affordable low-cost thermal cameras for monitoring of psychological states. Black et al. [4] compared state-of-the-art vision transformers (ViT) against traditional convolutional neural networks (CNNs) on data recorded with low-resolution thermal cameras. They found superior performance of ViT when classifying emotions, achieving 0.96 F1 score for 5 emotions (anger, happiness, neutral, sadness, surprise), confirming feasibility of low-cost hardware.

Lastly, some work explores subtle connections between different inner states that are difficult to discriminate, such as stress and CL. Bonyad et al. [5] showed correlation of the two states in airplane pilots, highlighting that elevated cognitive workload induced stress, manifesting in significant cooling across the nose, forehead, and cheeks, with the nasal region exhibiting the most rapid and pronounced temperature decline. These thermal changes were synchronized with increases in heart rate and subjective workload ratings. Overall thermal monitoring is becoming more accessible and an established CL estimation alternative to other modalities (e.g., wearables, RGB cameras, etc.), especially in challenging conditions (e.g., darkness).

3 Data

3.1 Data Collection

For the purpose of our experiment, we gathered data from 18 participants using various sensors. In this work, we will focus only on relevant data obtained by an affordable FLIR Lepton 3.5

camera, with resolution of 160x120 running at 8.7 frames per second.

Our participants underwent a battery of tests for inducing CL. Data collection was carried out in a controlled laboratory environment to ensure consistency across all participants. After filling out some initial questionnaires regarding individual's tiredness and focus levels, the calibration of various sensors used in the study was performed. The experiment itself was structured into three sequential blocks, each designed to induce CL through two different tasks offered at two difficulty levels. The first block featured standardized CL tasks – specifically, the N-back and Stroop tasks, which are widely used in cognitive research to engage working memory and executive attention [10, 12].

The second block introduced more ecologically valid memory tasks. The memory recall task involved displaying a list of words on a screen, after which participants had 30 seconds to recall and verbally report as many as possible. In the visual memory task, participants observed an image and were later asked to recall specific details.

The third and final block focused on ecological visual attention tasks. These included a visual discrepancy detection task and a line tracking task. In the discrepancy detection task, participants compared two images and identified visual differences. In the line tracking task, participants followed numbered lines from one side of the screen to the other and identified them.

Between these cognitive tasks, participants engaged in relaxation activities such as resting, passively viewing images, or listening to music, which served as baseline conditions and helped to balance their CL throughout the experiment. After each task and each relaxation period, participants completed the NASA Task Load Index (NASA-TLX) [7] and the Instantaneous Self-Assessment (ISA) [9] questionnaires to provide subjective evaluations of their cognitive and affective states.

The session concluded with the removal of all sensors, a debriefing session, and participant compensation. The entire procedure lasted approximately 60 minutes per participant, with around 40 minutes spent for active data collection and the rest used for setup, instructions, and debriefing.

3.2 Data Preprocessing

The raw data used in our analysis is illustrated in Figure 1. The first step in our preprocessing pipeline was windowing. Specifically, we divided each thermal video into consecutive 3-second windows with a 25% overlap. From each window, only the middle frame was selected for further analysis. This approach was based on the assumption that facial temperature changes driven by physiological responses such as blood flow occur gradually over several seconds rather than instantaneously. As such, a single representative frame from each interval was considered sufficient to capture meaningful thermal variation in 2.25-second steps.

The second step in preparing the data for subsequent machine learning experiments involved the extraction of features from thermal camera recordings. Prior research in this domain frequently utilizes average temperatures from distinct facial regions as input features, demonstrating that these regions can exhibit significant temperature differences associated with various affective states experienced by participants [3]. Motivated by these findings, we adopted a similar methodology to that proposed by Aristizabal-Tique et al. [3], and based our feature set on the average temperatures of four predefined regions of interest (ROIs): nose, forehead, left eye, and right eye.

The first step in obtaining the average temperatures for the selected ROIs involved applying a facial keypoint detector to extract the coordinates corresponding to each region in the thermal images. This process was carried out for the middle frame of every window of the thermal videos by passing it through a pretrained keypoint detection model [11]. The model, based on the widely adopted YOLOv5 architecture, was specifically trained on thermal images to enhance its performance for this modality. Following keypoint detection, we transitioned from working with raw thermal images to working with numerical temperature features, specifically the average temperatures computed for each region of interest. A more detailed explanation of this feature extraction process is provided in Section 4.1.

(a) Subject A. **(b) Subject B.**

Figure 1: Examples of raw thermal images.

At this stage, our dataset – where each row corresponded to a single video frame – contained a substantial number of missing values. These missing values were primarily due to limitations in keypoint detection, which stemmed from several factors. First, participants wore smart glasses during the experiment, which often obstructed the eye region and impaired the accuracy of the keypoint detector. Second, natural head movements, such as turning to the left or right, occasionally caused parts of the face to be occluded, preventing the detector from accurately identifying key facial landmarks. Given the impact of these issues on data quality, we chose to remove all rows containing missing values from further analysis. We excluded 31% of the data in this step. Use of smart glasses was not problematic only for keypoint detection, but also for feature calculation. The eye regions were partially obstructed by the glasses, thus preventing the thermal camera from capturing accurate temperature measurements in this area. Since we were unable to control for this effect, it is possible that it also posed an issue in classification.

Next, we performed label transformations to prepare the data for subsequent analysis. Initially, the dataset included multiple labels, each corresponding to one of the tasks described in Section 3.1. However, approximately 50% of the instances were labeled as "questionnaire", reflecting the periods during which participants completed self-report instruments such as NASA-TLX and ISA. These instances posed a challenge: filling out a questionnaire is neither a clear resting state nor a cognitively demanding task, making it difficult to accurately determine the level of CL involved. Since our primary interest lay in distinguishing between load and rest conditions, we opted to exclude all rows labeled as "questionnaire" from further analysis. In addition, we grouped the remaining labels into three broader categories: rest, low CL (corresponding to the easy versions of the tasks), and high CL (corresponding to the difficult versions).

Following some initial experiments, we chose to retain only the most "extreme" instances in terms of CL. Specifically, we excluded all data labeled as low CL, as this class exhibited substantial overlap with both the rest and high load conditions. In

particular, some tasks intended to induce low CL turned out to be unexpectedly difficult, effectively eliciting high CL, while others were so easy that it is questionable whether they imposed any cognitive demand at all.

To further emphasize the most distinct cognitive states, we also filtered the remaining data within each label interval. For intervals of instances labeled as rest, we retained only the final two-thirds of each interval, based on the assumption that participants would be most physiologically relaxed toward the end of each interval labeled rest. Immediately after completing a cognitively demanding task, the body may require some time to "cool down", during which residual physiological activity – such as elevated facial temperature – could still be present. By focusing on the latter portion of the interval, we aimed to capture a more accurate representation of the true resting state. Similarly, for instances labeled as high CL, we also retained only the final two-thirds of each interval, based on the assumption that CL tends to accumulate toward the end of a demanding task. This selection strategy was intended to maximize the contrast between rest and high load conditions by focusing on the time points most representative of those states.

4 Methodology

4.1 Calculating Features

As previously mentioned, we extracted features directly from the raw thermal images. Using the pretrained keypoint detector [11], we obtained coordinates for five facial keypoints, using which we then defined ROIs corresponding to specific facial areas for each 3-second window. ROIs were shaped as rectangles, positioned based on keypoint coordinates. Their size and placement were dynamically defined according to the distance between the eyes, reducing issues such as capturing inconsistent facial areas due to variations in distance from the camera or head movements. This approach was considered appropriate, because the study was conducted in a controlled laboratory environment with minimal variation in posture and setup. Additionally, a visual inspection of the extracted ROIs confirmed that they were well aligned.

Next, we computed the average pixel temperature for each ROI, as each pixel in a thermal image directly reflects a temperature value. This process yielded four primary features – one for each of the predefined ROIs (nose, forehead, left eye, and right eye). To capture relative temperature differences between these regions, we then computed the pairwise differences between all four average temperatures. This resulted in an additional six features, representing the thermal contrasts between different facial areas. Finally, to capture potential temporal trends in temperature changes, we introduced two additional temporal features. Specifically, for each 3-second window, we computed the temperature difference between the first and last frame for two key regions of interest: the nose and the forehead. These temporal features aimed to reflect short-term thermal dynamics that may be indicative of CL fluctuations. In total, this process resulted in 12 features per instance: 4 average ROI temperatures, 6 pairwise temperature differences, and 2 temporal difference features.

Finally, we applied personalized normalization to account for individual differences in baseline physiological responses. While there is considerable variability across participants, the variations within each individual are more informative for detecting changes in CL. To address this, we standardized all feature values using z-score normalization per participant, transforming each instance based on that individual's mean and standard deviation.

Table 1: Class distribution

Label	Count
Rest	1626
High Load	1548

This normalization helped reduce inter-subject variability while preserving intra-subject dynamics, enabling a more robust learning of patterns related to CL. Following this step, we proceeded with the machine learning experiments using the described set of features.

4.2 Experiments

After completing the data preparation steps outlined in Sections 3.2 and 4.1, we proceeded with the machine learning experiments. At this stage, the dataset consisted of two balanced classes: rest and high CL, as shown in Table 1. The models were trained on a total of 3174 instances, derived from 18 participants.

In our experiments, we employed a diverse set of machine learning models, including Random Forest (RF), Extreme Gradient Boosting (XGB), Stochastic Gradient Descent (SGD), k-Nearest Neighbors (KNN), and Light Gradient Boosting Machine (GBM). As a baseline, we included a majority classifier, which always predicted the most frequent class in the training data of each fold. Each model was trained and evaluated using its optimized hyperparameters, which were determined through a grid search strategy applied on training data on each Leave-One-Subject-Out (LOSO) iteration aimed at maximizing classification accuracy.

To ensure the robustness and generalizability of the results, we adopted a LOSO cross-validation approach, in which each participant served as a test subject exactly once while the remaining participants were used for training. This evaluation strategy is well-suited for personalized and physiological data, where inter-subject variability is high. To ensure a comprehensive evaluation of model performance, we did not rely solely on a single metric. Instead, we incorporated a range of evaluation metrics, including accuracy and F1-score. This multi-metric approach allowed us to better capture different aspects of model performance. The results of these experiments are presented in the subsequent section.

5 Results

As mentioned in the previous sections, we trained and evaluated a variety of models, and evaluated them using the LOSO. Summary of the results can be seen in Table 2, where both accuracy and F1-score are reported as averages across all subject folds, providing an overall measure of model performance and generalization performance.

The results indicate that the best-performing algorithm was SGD, achieving an accuracy of 0.64 ± 0.16, which represents a 0.13 improvement over the baseline majority classifier accuracy of 0.51 ± 0.00. In addition to its accuracy, SGD also achieved a high F1-score, suggesting that the model performs well in predicting both classes in a balanced manner. However, SGD also has the highest variance ($\pm$ 0.16), which indicates less stability across subjects. Overall, all evaluated models outperformed the majority class baseline. Moreover, the accuracy scores across all tested models were relatively similar, indicating consistent performance regardless of the specific algorithm used. Performance of GBM,

Table 2: Accuracy and F1-score of trained models compared to the majority class classifier

Classifier	Model Accuracy	Model F1	Majority Class Accuracy	Majority Class F1
RF	0.62 ± 0.13	0.62 ± 0.13	0.51 ± 0.00	0.34 ± 0.04
XGB	0.62 ± 0.14	0.62 ± 0.14	0.51 ± 0.00	0.34 ± 0.04
SGD	**0.64 ± 0.16**	**0.63 ± 0.16**	**0.51 ± 0.00**	**0.34 ± 0.04**
KNN	0.60 ± 0.10	0.60 ± 0.10	0.51 ± 0.00	0.34 ± 0.04
GBM	0.63 ± 0.10	0.60 ± 0.11	0.51 ± 0.00	0.34 ± 0.04

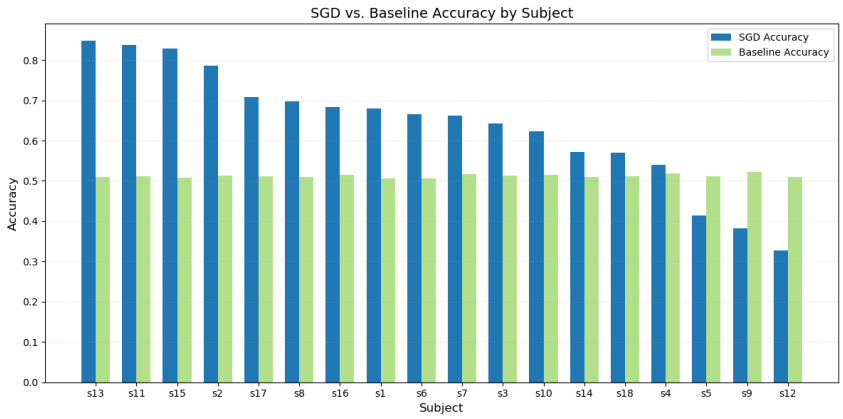

Figure 2: SGD vs. baseline majority classifier by subject.

RF and XGB was very similar, although somewhat behind the performance of the SGD.

Looking at per-subject results in more detail in Figure 2, we see that for most subjects, the SGD classifier outperformed the majority baseline classifier. SGD achieved its best performance on subjects 13, 11, and 15, with accuracies exceeding 0.80. There is also considerable variation across individuals, which aligns with the high variance reported in Table 2. This variability may indicate the presence of subject-specific patterns, label noise, or data that is inherently more challenging to learn.

6 Conclusion

This study demonstrates the potential of low-cost consumer thermal imaging as a viable, non-invasive method for estimating CL. By leveraging features extracted from key facial regions and applying various machine learning algorithms, we achieved promising results in distinguishing between rest and high load cognitive states. Among the tested models, SGD achieved the best average performance, though with notable inter-subject variability. These findings highlight both the strengths and current limitations of thermal-based CL estimation. While the results support the feasibility of using affordable thermal cameras in real-world applications, future work should explore strategies such as more sophisticated personalization to enhance generalization across individuals, deep learning, etc. This line of research points toward usefulness of cognitive monitoring in practical settings such as education, workplace safety, and adaptive user interfaces.

Acknowledgements

We sincerely thank our colleagues from Department of Intelligent Systems (Jožef Stefan Institute) for their assistance in data collection and preprocessing.

References

[1] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1, 3, 1–20.

[2] Muneeb Imtiaz Ahmad, Ingo Keller, David A Robb, and Katrin S Lohan. 2023. A framework to estimate cognitive load using physiological data. *Personal and Ubiquitous Computing*, 27, 6, 2027–2041. DOI: 10.1007/s00779-020-01455-7.

[3] Victor H. Aristizabal-Tique, Marcela Henao-Pérez, Diana Carolina López-Medina, Renato Zambrano-Cruz, and Gloria Díaz-Londoño. 2023. Facial thermal and blood perfusion patterns of human emotions: proof-of-concept. *Journal of Thermal Biology*, 112, 103464. DOI: https://doi.org/10.1016/j.jtherbio.2023.103464.

[4] James Thomas Black and Muhammad Zeeshan Shakir. 2025. Ai enabled facial emotion recognition using low-cost thermal cameras. *Computing&AI Connect*, 2, 1, 1–10.

[5] Amin Bonyad, Hamdi Ben Abdessalem, and Claude Frasson. 2025. Heat of the moment: exploring the influence of stress and workload on facial temperature dynamics. In *International Conference on Intelligent Tutoring Systems*. Springer, 181–193.

[6] Federica Gioia, Maria Antonietta Pascali, Alberto Greco, Sara Colantonio, and Enzo Pasquale Scilingo. 2021. Discriminating stress from cognitive load using contactless thermal imaging devices. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 608–611.

[7] Sandra G Hart and Lowell E Staveland. 1988. Development of nasa-tlx (task load index): results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[8] Stephanos Ioannou, Vittorio Gallese, and Arcangelo Merla. 2014. Thermal infrared imaging in psychophysiology: potentialities and limits. *Psychophysiology*, 51, 10, 951–963.

[9] CS Jordan and SD Brennen. 1992. Instantaneous self-assessment of workload technique (isa). *Defence Research Agency, Portsmouth*.

[10] Michael Kane, Andrew Conway, Timothy Miura, and Gregory Colflesh. 2007. Working memory, attention control, and the n-back task: a question of construct validity. *Journal of experimental psychology. Learning, memory, and cognition*, 33, (May 2007), 615–22. DOI: 10.1037/0278-7393.33.3.615.

[11] Askat Kuzdeuov, Dana Aubakirova, Darina Koishigarina, and Huseyin Atakan Varol. 2022. Tfw: annotated thermal faces in the wild dataset. *IEEE Transactions on Information Forensics and Security*, 17, 2084–2094. DOI: 10.1109/TIFS.2022.3177949.

[12] Michael P Milham, Kirk I Erickson, Marie T Banich, Arthur F Kramer, Andrew Webb, Tracey Wszalek, and Neal J Cohen. 2002. Attentional control in the aging brain: insights from an fmri study of the stroop task. *Brain and cognition*, 49, 3, 277–296.