International Journal of Social Robotics (2024) 16:981–997
https://doi.org/10.1007/s12369-023-01066-1

Multi-modal Affect Detection Using Thermal and Optical Imaging in a Gamified Robotic Exercise

Youssef Mohamed[1] · Arzu Güneysu[1,2] · Séverin Lemaignan[3] · Iolanda Leite[1]

Accepted: 14 September 2023 / Published online: 31 October 2023

Abstract
Affect recognition, or the ability to detect and interpret emotional states, has the potential to be a valuable tool in the field of healthcare. In particular, it can be useful in gamified therapy, which involves using gaming techniques to motivate and keep the engagement of patients in therapeutic activities. This study aims to examine the accuracy of machine learning models using thermal imaging and action unit data for affect classification in a gamified robot therapy scenario. A self-report survey and three machine learning models were used to assess emotions including frustration, boredom, and enjoyment in participants during different phases of the game. The results showed that the multimodal approach with the combination of thermal imaging and action units with LSTM model had the highest accuracy of 77% for emotion classification over a 7-s sliding window, while thermal imaging had the lowest standard deviation among participants. The results suggest that thermal imaging and action units can be effective in detecting affective states and might have the potential to be used in healthcare applications, such as gamified therapy, as a promising non-intrusive method for recognizing internal states.

Keywords Multi-modal affect recognition · Emotionally aware systems · Thermal imaging · Human–robot interaction · Frustration · Action units

1 Introduction

Affect recognition, or the ability to detect and interpret emotional states, has a number of potential applications in various fields, including gaming [1, 2] and healthcare [3]. In the realm of gaming, affect recognition can be used to enhance the gaming experience by allowing the game to respond to the emotional state of the player [4]. For example, a game may become more difficult or easier based on the player's level of frustration or may provide positive feedback when the player is experiencing enjoyment.

In order to induce the psychological state of flow in an individual, it is necessary to carefully regulate the level of challenge presented by the task or activity at hand. According to [5], an optimal level of challenge must be maintained, as deviations from this level can negatively impact the individual's experience. Specifically, a challenge level that exceeds the individual's current skillset may lead to feelings of anxiety or frustration, while a challenge level that falls below their skillset may result in boredom. Therefore, it is important to continuously monitor the individual's performance and adjust the task parameters accordingly in order to maintain the optimal level of challenge.

In the field of healthcare, affect recognition can also be useful in the context of gamified therapy, which involves the use of gaming techniques to motivate and engage patients in therapeutic activities [6]. By using affect recognition to monitor the patient's level of frustration and engagement, the therapist can adjust the therapy as needed to keep the patient motivated and engaged. This can be especially important in situations where the patient is participating in the therapy remotely, as it can be more difficult for the therapist to gauge the patient's emotional state without non-intrusive methods of affect recognition [7] or in home-based therapy where the system needs to adapt itself automatically. Overall, the use of affect recognition in both gaming and healthcare applications can help to improve the effectiveness of these interventions by providing real-time feedback on the emotional state of the participant.

✉ Youssef Mohamed
ymo@kth.se

[1] EECS, Division of Robotics, Perception, and Learning, KTH Royal Institute of Technology, Stockholm, Sweden

[2] Digital Futures, Stockholm, Sweden

[3] PAL Robotics, Barcelona, Spain

Emotionally aware systems might have the potential to improve the levels of trust and acceptance of robots in healthcare settings, thus making them a more beneficial tool for vulnerable populations. Affective computing strategies can be employed to enable robots to better recognize and interpret human emotions, as well as respond in an appropriate manner [8]. This can lead to an improved, meaningful relationship between robot and user and create a strong connection that motivates and encourages engagement in robotic technologies [9]. As such, emotionally aware robots have the potential to be more trustworthy, thus making them more acceptable in healthcare environments.

It has been found that some affective states when interacting with patients are more prominent than others, like frustration, enjoyment, and boredom [1, 2, 10]. Upon detecting these affects the robot will be able to adapt to the patient's needs and capabilities. Boredom and frustration can be useful to consider when designing emotionally aware robots because they are common emotions that humans experience in various situations. For example, a person using a robotic device for exercise therapy may become bored if the device does not provide enough variety or challenge, or if the therapy sessions are too repetitive. Similarly, a person may become frustrated if the device is difficult to use or if it does not provide the desired results. Incorporating the ability to recognize and respond to these emotions in robots could potentially improve the user's experience by making the technology more engaging and effective. For example, a robot that is able to detect when a user is becoming bored or frustrated could adapt its behavior to provide more variety or challenge or offer suggestions for ways to make the therapy more enjoyable. This could help to maintain the user's motivation and engagement, and ultimately make the technology more beneficial and trustworthy.

Current methods for extracting social signals in robotics include using facial landmarks, action units, and pose estimation. These methods have been used in a variety of applications, such as in detecting genuine facial expressions [11] and in educational settings to detect engagement through body posture [12]. Nonetheless, inferring affective states and understanding those signals can be skewed, biased, and/or subjective [13, 14].

The interpretation of affect can vary greatly from one person to another. Thus, several sensors have been introduced to detect those affective states using different physiological signals, including electrocardiography, electromyography, skin conductance, and body temperature [15]. However, these sensors are usually intrusive and can affect the patient's behaviour [7], making them unsuitable for real-world scenarios. Intrusive sensors face several limitations for human–robot interaction that cameras can mitigate. Intrusive sensors can deteriorate over time due to constant skin contact or frequent use, limiting their longevity. In contrast, cameras have no direct contact with the human, allowing for potentially indefinite deployment. Furthermore, intrusive sensors disrupt natural social interactions and flow, as they require users to attach multiple sensors to their body, which can be cumbersome and obtrusive. This obtrusiveness is especially problematic for dynamic, multi-person social scenarios. Cameras, as non-contact sensors, seamlessly integrate into social contexts and require little to no calibration from the user. This seamlessness and ease of use is essential for natural human–robot interaction. In formal or informal social settings, people expect to simply engage with the robot immediately, rather than taking minutes to attach numerous sensors. For complex social situations involving multiple participants, intrusive sensors become even more disruptive, while cameras can continue operating unobtrusively.

Thermal imaging has been gaining attention in recent years for its potential to detect internal states like stress [16] and cognitive load [17, 18]. This is due to the automatic reactions of the sympathetic nervous system, which are reflected in facial temperature [19–21]. Thermal cameras are becoming more accurate and affordable, making them a promising tool for detecting internal states. There are four main states that determine pivoting points during an interaction, where the robot will have to adapt to the affective states of the participants: frustration, boredom, enjoyment, and neutrality [1, 2, 10]. Therefore, in this study, we utilize the use of thermal imaging and facial expressions to detect four states: frustration, boredom, enjoyment, and baseline during a gamified exercise of playing PacMan with tangible robots which is iteratively designed for upperlimb rehabilitation [22].

In this work, we aim to classify four affective states: frustration, boredom, enjoyment, and baseline, experienced by participants while playing a gamified exercise. To accomplish this, we collected data from 30 participants who engaged in the exercise. We then used this data to train a Long Short-Term Memory (LSTM) model and a Gaussian Naive Bayes (GNB) model. The LSTM model is used to capture the temporal aspect of the data, while the GNB model served as a simple probabilistic comparison. To ensure the accuracy of the models and prevent over-fitting, we employed a leave-one-out approach for testing.

2 Related Work

2.1 Multi-modal Affect Detection

Multi-modal approaches focus on detecting multiple affective states; however, one of the most commonly experienced affective states in HRI is frustration [23]. Therefore, several approaches have focused on detecting it. For example, in [24] the authors used a set of sensors, including skin conductance, pupil trackers, posture, and pressure sensors to predict frustration. The authors recruited 24 participants to interact with a tutoring virtual agent while doing a “Towers of Hanoi” activity. The best-performing model was a Gaussian Process model which reached an accuracy of 79%. Taylor et al. [25] have used a similar approach by making use of three wearable sensors to detect frustration: Electro-Dermal Activity (EDA), Heat Flux, and Skin Temperature. The participants were instructed to play a modified version of the game "Breakout", on which the researchers had introduced some latency to induce frustration. Naïve Bayesian models were trained to classify frustration, reaching an accuracy of 80%. In both of these works, although the models have reached high accuracies, the sensors used can be intrusive and impractical in more socially dynamic environments.

Hence, other approaches such as in [26] used the RGB camera to detect Facial Action Units (FAUs) during a playground of physics in the wild in a classroom environment, where participants had to apply basic physics principles to solve a puzzle. The authors have used both face-only and interaction-only features, and both have reached ROC AUC above chance to detect: boredom, confusion, frustration, delight, and engagement as a multi-class classifier, reaching even higher AUC values when turned into a binary classifier with each state. The approach used is similar to this study; nonetheless, a learning environment is still constrained and could have had various social and psychological impacts from a closed classroom. In addition [26] depends on the RGB facial features extracted by FACET, without the use of any other automatically extracted modality.

Other affective states like boredom and enjoyment are also considered critical to detect in HRI, and healthcare environments as they are considered pivotal moments during the interaction [1]. In [27] the authors created a Random Forests (RF) binary classifier that can detect boredom. The features used were from both the body language and the gaze data, then both these modalities are considered by a human coder to score, whether the person was bored or not based on the naive definition of boredom, for both gaze and skeletal data. Then, the model inputs are the scores given in 5-s intervals of the video data collected. The created model was able to detect boredom states with an F1 score of 78%, compared to the baseline state of 43%.

When looking into enjoyment as a state, it can be seen that several synonyms are used interchangeably in the literature like happiness or joy. In this work, we use the word enjoyment because participants are doing a task for a limited time that they can enjoy, but we cannot make larger claims about their mental state or a generalization if they are happy or joyful. There are numerous works on detecting these states, as they are considered one of the 6 basic emotions [28–30]. Nevertheless, one of the most influential multimodal approaches used in [31], the authors have used prosody from audio and facial expressions from the video as inputs to their models. Several models were tested using WEKA and the best performing model was JRip. The model detected happiness, sadness, neutrality, and surprise with an accuracy of 75.5%, after combining both data types.

2.2 Thermal Affect Detection

Vision-based cameras are commonly used for the extraction of action units and body movements. For example, [32] used a Microsoft Kinect for six basic emotion predictions. A unimodal neural network was trained on both the facial expression and body movement streams using late fusion. 93% was the accuracy achieved by the neural network.

Although the use of RGB cameras can lead to high-performing models, these cameras are dependent on the lighting conditions of the recorded dataset and other environmental conditions. Self-reported measures and conflicting facial expression labels are other factors that these models can be heavily affected by [33].

Alternatively, thermal cameras use far infrared to measure the radiation emitted by warm objects, which is independent of reflected light [34]. Therefore, thermal imaging can be used to overcome the limitations of RGB cameras, as the thermal spectrum is not affected by the presence of light and can record objective measures such as changes in skin temperature [35]. It has been established in the literature that stress and cognitive load have apparent effects on skin temperature [36–40], motivating the use of thermal imaging for affective state detection in HRI scenarios. In [41], a thermal camera was mounted on a Meka robot to measure facial temperature variations while playing a card-based quiz game with the robot. The authors tested different environmental setups with the positioning of the robot. They concluded that significant effects can be seen on the nose temperature of the participant when the robot is positioned closer to their personal space, causing a higher stress response.

In a previous study [42], we examined the utility of thermal imaging and facial expressions as indicators of frustration in a laboratory setting. Participants were subjected to two types of frustration: cognitive load-induced frustration and failure-induced frustration; the latter occurs when a person is unable to overcome the cause of the failure [43]. We found that thermal imaging alone was effective at detecting frustration in both types of frustration, with similar accuracy to models trained on RGB features. Specifically, the highest accuracy for thermal data was achieved using three facial regions of interest: the nose, forehead, and lower lip. Our model reached an accuracy of 81% using RGB features, 64% using only thermal features, 55% using EDA, and 74% using all modalities. Our findings suggest that thermal imaging may be a valuable tool for detecting frustration in a controlled setting.

3 Methodology

3.1 Data Collection

3.1.1 Robotic Platform and Tangible Gamified Exercise

The Cellulo platform [44] is a system that allows users to interact with small palm-sized robots by moving them on printed paper sheets. These robots are equipped with illuminated capacitive touch buttons, and can be connected to a mobile device via Bluetooth. The paper sheets are overlaid with a microdot pattern (barely visible to the naked eye) that enables accurate (x, y, θ) self-localization of the robot with sub-millimeter precision.

Specific active zones of the printed map can be associated to pre-defined robot behaviors, allowing for the creation of mobile, physical game elements that can act as autonomous agents or input devices. The combination of paper sheets, robots with specific interaction modalities and behaviors, and mobile device software enables the development of unique game designs that incorporate physical exercise and interactive elements.

The tangible gamified exercise that we used in this study was previously co-designed with the stakeholders and used as a platform for children with special needs [45], patients going through upper limb rehabilitation [46], and for healthy aging [47].

The game is inspired by the classic *Pacman* game, and the objective is for the player to collect all six apples on the map as quickly as possible, while avoiding being caught by autonomous 'ghost' robots or crashing into the walls. The player manipulates the 'Pacman' robot by physically moving it around a map and collecting apples, which are represented by lights on the robot. In some configurations, collisions with walls result in a penalty, requiring the player to recollect the most recently acquired apple. The round ends when the player collects all six apples, at which point the 'ghost' robots return to their starting positions and a new round can begin.

The game allows for the customization of the number of autonomous agents, or 'ghost' robots, that can be set to chase the player, with options for one or two agents. Additionally, the speed of these agents can be adjusted to suit the user's preference, with measurements in millimeters per second (mm/s). The platform also includes an option to enable a *spin* rule, where fruits can only be collected by spinning the robot by 90° or 180° while hovering the apples, and a *wall crash penalty* rule which causes the player to lose the last collected fruit upon collision with a wall. To complete the immersive experience, haptic feedback can be enabled to assist the player with informative assistance when crashing into a wall. Overall, these customizable elements provide the users with a challenging and immersive experience.

Before collecting data, several pilot experiments were conducted to fine-tune the programmable behaviors of the robots, including factors such as speed, implementation of the *spin* rule, and the number of ghosts, in order to elicit the intended affective states in the majority of participants. The optimized parameters determined from these initial experiments were then used in the main setup with confidence.

3.1.2 Experimental Design

In this experiment, 30 participants aged between 19 and 32 (47% female and 53% male) played a modified version of the gamified exercise using the Cellulo platform. One participant was excluded from the experiment due to technical issues leaving 29 participants. Before the experiment, all participants signed a written informed consent form and got a compensation of 100 Swedish kronor (about 9 euros).

During data collection, each participant played a series of games with 4 different modes, each was designed to induce one of the affective states. The first phase was *familiarization* which is intended for the participants to get used to the game and for their face temperature to reach baseline values in the room. The familiarization phase included playing two rounds of the game to get familiar with the dynamics, game rules, and robotic manipulation. In this phase only one robot was following the user, with a speed of 70 mm/s. The second game round included showcasing the *spin* rule, where the participants had to spin the robot by 45° to collect the apple with the same game configuration.

It was followed by a 10 min period, during which the participants filled up the questionnaires and stayed silently in the room.

The second phase was designed to induce *enjoyment*. The participants had to rotate the robot by 90° to collect the points in all rounds, with starting speed of 100 mm/s and after every three rounds, the speed of the robot would increase by 20 mm/s. 2 ghosts would be following the user's robot from the start of the phase.

Previous studies with patients and healthy participants showed that the changes in speed and rotation angle during the second phase of the game make the game more challenging [22, 48] which is also observed as making the game more enjoyable for the participant if the speed does not increase drastically. By increasing the speed of the robot after every three rounds, the game may become more dynamic and require the participant to think and react quickly in order to collect the points. Additionally, the change in angle may add an additional layer of difficulty, as the participant must adjust their movements to account for the new direction in which the robot is moving. These changes make the game more engaging and stimulating, leading to higher enjoyment for the participant.

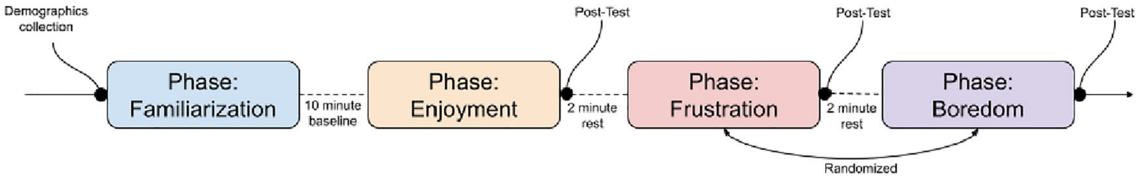

Fig. 1 Description of the study design, consisting of four phases: *baseline*, *enjoyment*, *frustration*, and *boredom*, followed by a questionnaire indicated as "Q" only after the 3 main provoked states

Table 1 Questionnaire used after each game phases

Questions
Mental demand: how mentally demanding were the rounds?
Physical demand: how physically demanding were the rounds?
Temporal demand: how hurried or rushed was the pace of the rounds?
Performance: how successful were you in accomplishing what you were asked to do?
Effort: how hard did you have to work to accomplish your level of performance?
I would describe these rounds as very interesting
How insecure, discouraged, irritated, and stressed were you?
I thought the rounds were quite enjoyable
I thought these rounds were boring
These rounds were fun to do
How frustrated and annoyed were you during these rounds?

The third phase was *frustration* and we hypothesized that frustration would be induced through very hard game rules which might require excessive effort from the user [46]. Therefore we introduced the *spin* rule with 180° rotation to collect the points, the *penalty* rule where the participant would lose points when they hit the wall, and the robot speed was set to 150 mm/s. When the speed of the robot is increased while being chased by two ghosts, the game becomes more difficult to beat. In addition, when the participant has collected one point with great difficulty, it is easy to lose that point if the robot touches the wall on the map. This is aligned with the definition of frustration, where the participant is trying to reach a goal but constantly failing at achieving it [49].

The fourth phase was designed to induce *boredom*: only one ghost was following the users 'Pacman' robot, and the speed of the ghost was reduced to 50 mm/s (see Fig. 1). The noticeable decrease in speed, and the lack of challenge with repetition of the rounds, would lead the participants to not get much stimulation while playing the game.

The order of the frustration and boredom phases was randomized and balanced among participants, while the enjoyment phase was always presented second in order to avoid the potential for participants to undergo the effect of boredom or frustration even though we introduce an enjoying condition or feeling bored due to the overall duration of the experiment rather than the specific parameters of the individual tasks.

The participants were asked to complete a questionnaire after each phase of the study followed by a 2 min rest. The questionnaire (Table 1) consists of eleven questions that were designed to evaluate their levels of frustration (two questions), boredom (one question), and enjoyment (three questions) on 5-Point Likert scales. The questionnaire was a combination of NASA-TLX [50] and Intrinsic Motivation Inventory (IMI) [51].

3.1.3 Self-Reports

Although we designed the game configurations of each game phase to induce enjoyment, boredom, and frustration, each user might have different preferences and affective responses for these phases. Therefore we introduced the above-mentioned questionnaire after each phase to understand the self-perceived affective states of the participants after the game-play of these phases.

In Fig. 2 you can see the self-reported scores of affective states in each phase of the game. The first plot shows the enjoyment scores where we have the highest number achieved in the Enjoyment Phase (E) that we designed with an average score of 4.1 out of 5. It shows that the enjoyment phase was successful in inducing enjoyment. The second plot presents self-reported frustration scores and the highest score (average is 3.2 out of 5) belongs to the Frustration Phase (F) that we designed to introduce frustration through hard game rules. The third plot shows the self-reported boredom scores and the highest score (average score of 3 out of 5) belongs to the Boredom Phase (B) that we designed with very simple and repetitive game configurations.

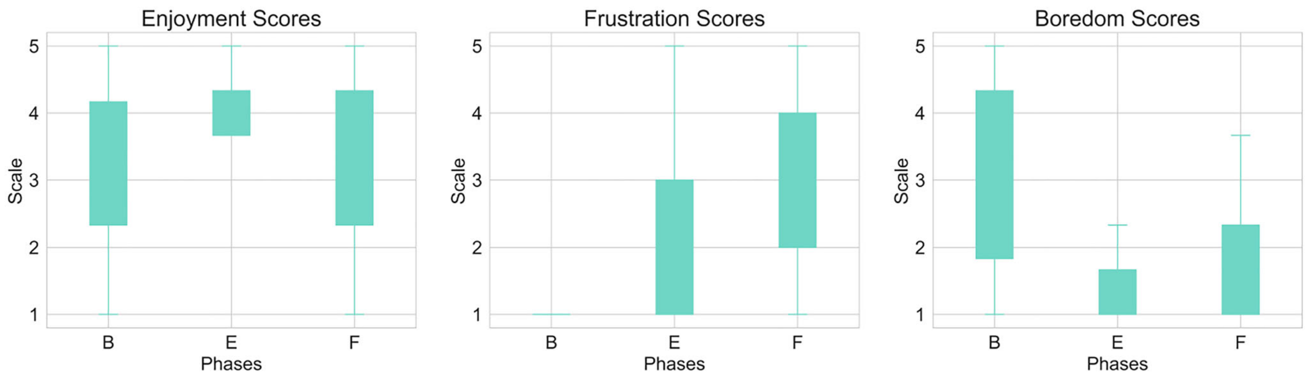

Fig. 2 Box plots of self-reported affective state scores for each of the game phases: boredom phase (B), enjoyment phase (E), and frustration phase (F)

Table 2 Results of the ANOVA and posthoc testing for 29 participants self-reports while trying out the game, the results show the *p*-value and the mean difference of each of the scales compared to the phases which were meant to induce the affects stated

Phases	Scales					
	Enjoyment		Frustration		Boredom	
	p-value	Mean-diff	*p*-value	Mean-diff	*p*-value	Mean-diff
Boredom-enjoyment	0.001	0.89	0.03	0.54	0.001	– 1.3
Boredom-frustration	0.6	0.89	0.001	1.4	0.001	– 0.98
Enjoyment-frustration	0.01	– 0.68	0.001	0.86	0.3	0.32

Each two phases scales are compared to each other

Furthermore, we have conducted an ANOVA with a Tukey post-hock to identify the phases which have significant differences between participants.

On the frustration scale, all three phases were significantly different while on the boredom scale boredom was significantly different from enjoyment and frustration. Furthermore, on the Enjoyment scale, enjoyment is significantly different than boredom and frustration (Please see Table 2). All these results confirm that the designed game phases induces the expected affective states relying on the self-reported perceived affective states of each user.

3.2 System Implementation

The system architecture (Fig. 3) was composed of two cameras jointly calibrated (thermal IR camera: `Optris PI 640`[1] and RGB-D camera: `RealSense D435`[2]) and `Cellulo`[3] [44] robots. All of the mentioned components were synchronized in real-time using the Robotic Operating System (ROS). In addition, OpenCV was used for image processing (cropping, creating ROI) and camera calibration.

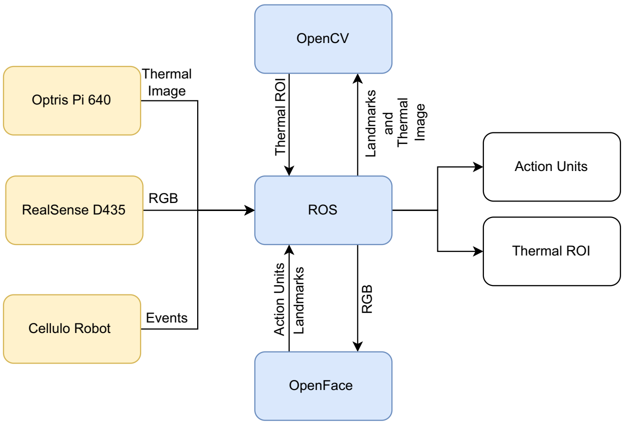

Fig. 3 System architecture

3.2.1 Extracted Features

The thermal imaging and RGB data were combined and synchronized. Following data collection, the data was tagged based on the self-reports for each phase, but only if the participant's score exceeded 2 during each of the phases in the self-reported emotion. The rationale behind using the cutoff score of 2 in our study was twofold: Firstly, we aimed to capture participants' emotional responses during the most engaging and challenging segments of the interaction. We hypothesized that these segments would elicit stronger emotional responses, and therefore, by concentrating on phases

[1] https://www.optris.global/thermal-imager-optris-pi-640.

[2] https://www.intelrealsense.com/depth-camera-d435/.

[3] https://www.epfl.ch/labs/chili/index-html/research/cellulo/.

Table 3 Extracted features for each modality

Modality	Features	ROIs/AU
Thermal	ROIs temperature average	Nose, Forehead, Cheek, Lowerlip
	ROIs temperature change	
	ROIs temperature maximum	
	ROIs temperature minimum	
RGB	AU intensity average	1, 2, 4, 5, 6
	AU intensity change	7, 9, 10, 12, 14
	AU maximum intensity	15, 17, 20, 23, 25
	AU minimum intensity	26, 28, 45

In total, $n = 88$ values are computed

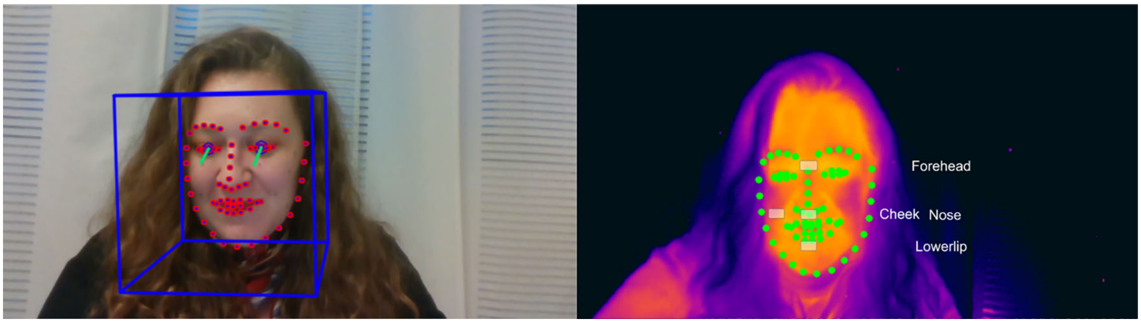

Fig. 4 Face landmarks and action units extracted using OpenFace on the left and thermal regions on the right image

where participants self-reported higher levels of emotion, we expected to obtain labels that more accurately represented the participants' emotional experiences during the key interactive parts of the study. Secondly, self-reported emotional scores of 2 or lower typically indicate relatively neutral or low arousal emotional states. These subtle emotional signals can be challenging to detect using computer vision and thermal imaging techniques. Consequently, by excluding phases with low self-reported emotion, we intended to label data where participants demonstrated clear and detectable emotional responses. This approach aimed to ensure that the labels we generate are meaningful and useful for training our automated multi-modal emotion detection models.

Six participants scored 2 or less in the frustration phase, 1 in the enjoyment phase, and 5 in the boredom phase.

The features extracted from the data are shown in Table 3. The features were extracted using both OpenCV and OpenFace as mentioned in Sect. 3.2 to extract regions of interest (ROIs) including nose, forehead, cheek, and lower lip (Fig. 4). Each of the regions is the average temperature within the region surrounding the facial landmark corresponding to that region [18].

Thermal data was collected at a rate of 15 frames per second (fps). RGB camera data (action units) was collected at the same frequency. The features for the thermal data were computed for all four facial ROIs: nose, forehead, cheek and lower lip. As for the action units extracted, they corresponded to the Facial Action Coding System (FACS) [52]: 1 (inner brow raiser), 2 (outer brow raiser), 4 (brow lowerer), 5 (upper lid raiser), 6 (cheek raiser), 7 (lid tightener), 9 (nose wrinkler), 10 (upper lip raiser), 12 (lip corner puller), 14 (dimpler), 15 (lip corner depressor), 17 (chin raiser), 20 (lip stretcher), 23 (lip tightener), 25 (lips part), 26 (jaw drop), 28 (lip suck), and 45 (blink).

For each data point i, we collect a total of $n = 88$ measurements $M_i = [m_0, \dots, m_n]$ (Table 3), as well as a label l_i. Feature extraction for classification is performed through a sliding window of predefined length ($L = 3.5s$ or $7s$) [42, 53] and hop length $h = 0.5s$. For every window (i.e. each instances), we compute a feature vector $X_j = [x_0, \dots, x_n]$ (used for training and testing) by taking the mean, the delta (difference between the starting and ending value within the window), the maximum and the minimum values over all data points in that window, for each measurement in M. The label Y_j of that instance is given by the most common label l within that window.

While we maintain the window length used in our previous work, we choose overlapping windows because this better reflects how a real-world system would operate (e.g. to provide continuous estimations). This process is illustrated in Fig. 5. In Table 4 the total number of instances obtained for each window length is shown.

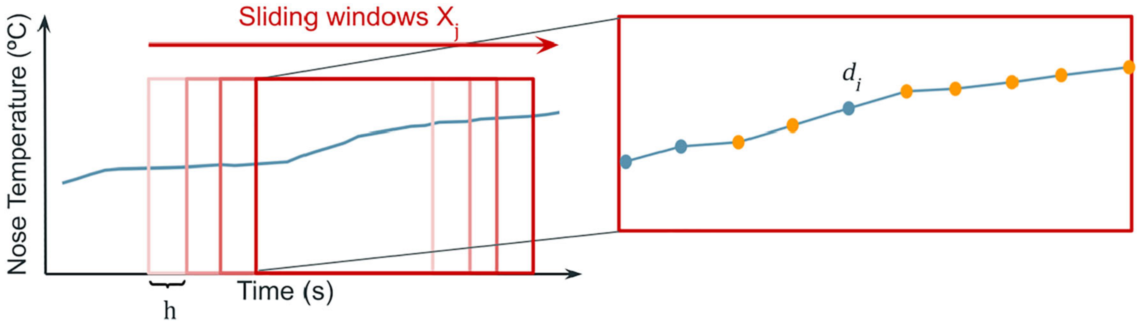

Fig. 5 Schematic view of feature extraction. Measurement m_i (in this case, nose temperature), is associated with a set of data points d, each labelled according to the self-reported state of the participant. One instance X_j for training is composed of features that are calculated over the set of data points d, such as average nose temperature. Y_j, the label of this training instance, is given by the most common label (in this example, $Y_j = F$)

Table 4 Total number of instances used for training in all four states baseline, enjoyment, boredom, and frustration

Window (s)	No. of instances			
	Baseline	Enjoyment	Boredom	Frustration
3.5	22,642	15,412	16,602	15,248
7	22,427	15,287	16,392	15,094

3.3 Models

The problem we address in this work is the classification of affective states in human subjects, based on thermal and visual features. We chose to represent this problem as a multiclass classification task, where the goal was to predict one of four classes: Frustrated, Bored, Enjoyed, or Baseline.

In order to compare the performance of different models, two were selected for analysis: a Long Short-Term Memory (LSTM) model and a Gaussian Naive Bayes (GNB) model. The LSTM model is designed to take into account temporal dependencies [54–56], while the GNB model is a simple probabilistic classifier [57]. The performance of these two models will be evaluated and compared in order to determine which one is more effective for the given task.

The architecture of our LSTM model is illustrated in Fig. 6. During training, we applied dropout regularization to prevent overfitting.

LSTM hyperparameters included the following:

- LSTM networks: *1*
- Dropout rate: *0.6*
- Batch size: *128*
- Hidden layers: *2*
- Learning rate: *0.001*
- Number of epochs: *70*

For comparison, we choose to use the GNB model from the scikit-learn library. This model is a supervised learning algorithm that utilizes Bayes' theorem to make predictions. The Gaussian Naive Bayes model assumes that the data is distributed normally and that the features are independent of each other. The model was trained on the features extracted from the thermal and visual data, and was used to predict the four affective state. The hyperparameters included variable smoothing of: $1e{-}09$

The choice was based on testing multiple machine learning algorithms: Random Forest Classifier (RFC), Support Vector Machine (SVM) and K-nearest Neighbor (KNN). The models testing was done using a Grid Search Cross Validation (GSCV) algorithm, which tested each of the models on a range of hyper-parameters, outputting the ones with the highest accuracy. None of the tested algorithms showed accuracies above chance, only the GNB showed higher performance. Furthermore, the use of GNB has been proven to perform well for affect detection using facial expressions [58, 59].

3.3.1 Preprocessing

The input to the models consisted of the thermal and visual features mentioned in the paper. The features were prepro-

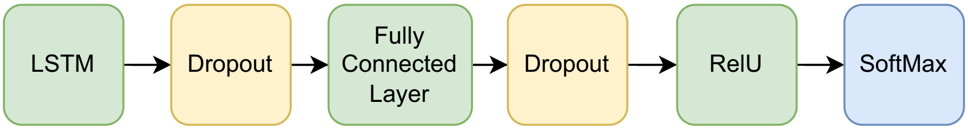

Fig. 6 The architecture of the LSTM model consisting of a layer of LSTM, 2 Dropout layers, a fully connected layer, relU, and SoftMax as a final layer

cessed by being filtered (values $25 < T < 39$) and then standardized. The output of the models was a prediction of one of the four classes, which was made at regular intervals of either 3.5 or 7 s during training and testing.

We used a leave-one-group-out cross-validation approach. Specifically, we divided our dataset into 29 groups, with each group containing the data from one participant. We then trained our models on data from 28 groups and evaluated their performance in the remaining group. We repeated this process for each group in our dataset, resulting in 29 total evaluations. We used the same train-test set-up for all models and repeated the process for each participant with different random seeds to ensure the robustness of our results.

We opted to use the micro-average F1 score as our primary evaluation metric, primarily due to the balanced nature of our dataset. As shown in Table 4, each class has a roughly equal number of instances, which makes the F1 score a more suitable metric than accuracy alone. Accuracy can be misleading in cases where the dataset has imbalanced class distributions, as it can yield high values even if the classifier performs poorly on the minority class. The F1 score, on the other hand, provides a more comprehensive evaluation by considering both precision and recall, which are crucial to understanding the true performance of a classifier. By using the micro-average F1 score, we ensure that each instance contributes equally to the overall performance measure, making it a suitable choice for our balanced dataset.

4 Results

This section reports the results of both accuracy and F1 scores for each of the three model variations (thermal, AUs and thermal + AUs) to classify the affective states of frustration, enjoyment, boredom and baseline.

4.1 Models Performance

4.1.1 LSTM

In Table 5, the accuracy and the F1 percentage are shown, including the standard deviation across participants tested using the leave-one-out approach. It can be seen that in all three model variations, the 7-s window has higher accuracy compared to the 3.5-s window. In addition, thermal imaging on its own being a lower accuracy, and thermal + AUs being the highest accuracy of 77% in the 7-s window. Furthermore, thermal imaging shows the lowest standard deviation between participants while using the leave-one-out method being only 5% compared to 12–13% for the AU and thermal + AUs

The average of the predicted values for each of the cross-validation sets was considered to create a confusion matrix that would describe the performance of the model (See Fig. 7). The normalized confusion matrix presented demonstrates the performance of an LSTM model in classifying four affective states: Baseline, Enjoyment, Boredom, and Frustration. The diagonal values indicate the proportion of correct predictions for each class. The model achieved the following classification accuracies: Baseline (0.89), Enjoyment (0.72), Boredom (0.77), and Frustration (0.69). showing that the highest predicted affect is baseline and the lowest is frustration.

4.1.2 GNB

It can be seen in Table 6 that the Thermal + Action Units modality performs the best, achieving the highest accuracy and F1 scores. Specifically, when using a window size of 7 s, the model achieved an accuracy of 35% and an F1 score of 38% with a standard deviation of 17%. When using a window size of 3.5 s, the model achieved an accuracy of 34% and an F1 score of 37% with a standard deviation of 15%. This indicates that the model performed well in classifying the four classes, and that the performance improves as the window size increases from 3.5 to 7 s.

On the other hand, the Thermal and Action Units modalities alone performed relatively worse in comparison, with a lower accuracy and F1 scores. The Thermal modality achieved an accuracy of 28% and an F1 score of 30% with a standard deviation of 18% when using a window size of 3.5 s and an accuracy of 28% and an F1 score of 32% with a standard deviation of 17% when using a window size of 7 s. The Action Units modality achieved an accuracy of 29% and an F1 score of 33% with a standard deviation of 15% when using a window size of 3.5 s and an accuracy of 30% and an F1 score of 34% with a standard deviation of 16% when using a window size of 7 s.

Table 5 LSTM model performance for each modality and for each window size for classifying 4 classes (random is 25%)

Modality	Thermal		Action units		Thermal + action units	
Window size (s)	3.5	7	3.5	7	3.5	7
Accuracy (%)	53 ± 6	58 ± 5	68 ± 11	70 ± 12	75 ± 12	77 ± 13
F1 (%)	53 ± 6	58 ± 5	67 ± 12	69 ± 14	75 ± 12	78 ± 13

Showing the accuracy and the F1 score of each model, including the standard deviation across participants using the leave-one-out approach

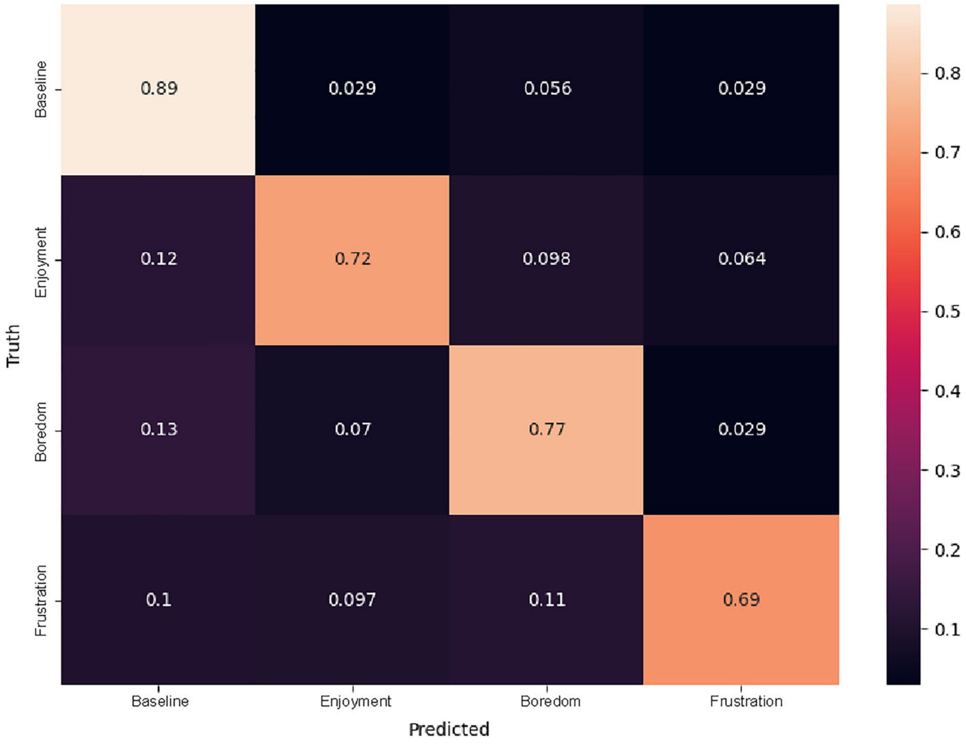

Fig. 7 Heatmap of the confusion matrix of the thermal + action units LSTM model in the 7 s window, representing the accuracy of each class

Table 6 GNB model performance for each modality and for each window size for classifying 4 classes (random is 25%)

Modality	Thermal		Action units		Thermal + action units	
Window size (s)	3.5	7	3.5	7	3.5	7
Accuracy (%)	28 ± 17	28 ± 17	29 ± 15	30 ± 15	34 ± 17	35 ± 17
F1 (%)	30 ± 18	32 ± 17	33 ± 15	34 ± 16	37 ± 15	38 ± 17

Showing the accuracy and the F1 score of each model, including the standard deviation across participants using the leave-one-out approach

4.2 Feature Importance

Feature selection is a crucial aspect of machine learning model development, as it involves identifying the most relevant and informative features to include in the model. One way to assess the importance of a given feature is through the use of permutation importance, as described in [60]. This method involves evaluating the effect on the model's performance, as measured by a metric such as the F1 score when the values of a single feature are permuted. If permuting the values of a particular feature significantly impacts the model's performance, it can be concluded that the feature is an important contributor to the model's prediction. In our study, we computed the permutation importance of the features in the best model and measured the increase in the F1 score after permuting the values of each feature. This allowed us to understand the relative importance of each feature in relation to the model's prediction.

The results, depicted in Fig. 8, demonstrate that the RGB data scores generally had higher permutation importance scores in comparison to the thermal data scores. Specifically, min_AU26 possessed the highest permutation importance score of 20.91, followed by max_AU12 with a score of 19.18 and delta_AU23 with a score of 18.74. On the other hand, the thermal data scores had lower permutation importance scores in comparison to the RGB data scores. For instance, min_Forehead had a permutation importance score of 17.51, delta_Lowerlip had a score of 10.95, and min_Cheek had a score of 8.32.

This implies that the model is relying more heavily on the RGB data for its predictions. It could also indicate that the RGB data is of higher quality or has been better preprocessed for the model's use. These findings might suggest that the model could potentially be improved by focusing on the quality or preprocessing of the thermal data, or by incorporating additional thermal data features that may be more informative.

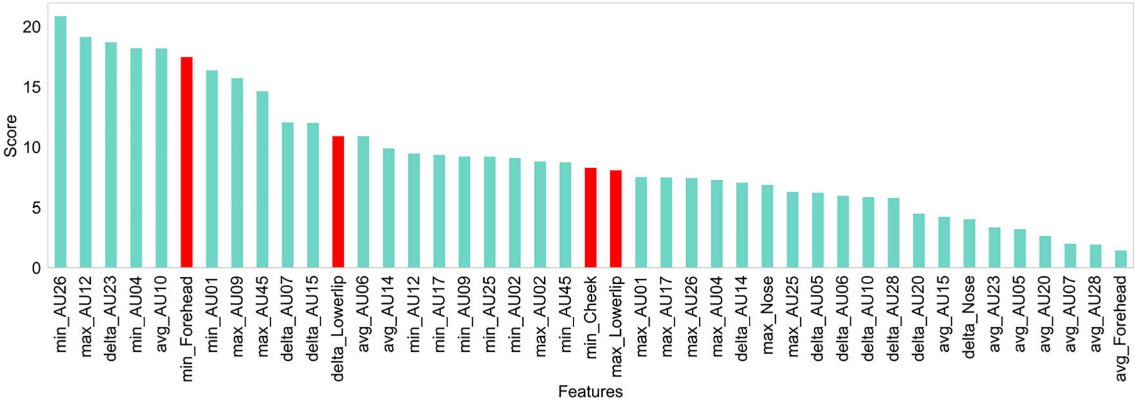

Fig. 8 Feature Importance for F1 Score for 43 Features. The permutation importance of each feature is shown on the y-axis, with the features sorted by their importance on the x-axis, red represents thermal features and blue is action units

5 Discussion

The self-reports used a 5-point scale to measure participants' levels of frustration, boredom, and enjoyment during the phases of the game. The data collected indicates that the game effectively induced the three affective states, as shown by significant differences in the levels of enjoyment, frustration, and boredom between the boredom and enjoyment phases. This suggests that the game was successful in creating distinct emotional experiences for the participants.

The performance of three different machine learning models were evaluated for classifying affective states of participants based on their thermal imaging and action unit data. The results demonstrate that the 7-s window yielded higher accuracy compared to the 3.5-s window for all three modalities. Thermal imaging alone had the lowest accuracy, while the combination of thermal imaging and action units had the highest accuracy of 77% in the 7-s window. Furthermore, thermal imaging had the lowest standard deviation among participants, while the combination of thermal imaging and action units had the highest standard deviation.

Furthermore, in examining the performance of the LSTM model in each class (Fig. 7), the confusion matrix reveals that the model is most successful in identifying the Baseline state, with an accuracy of 0.89. This indicates that the model is effective in distinguishing between neutral affect and the other affective states. Additionally, there is a notable degree of misclassification between Boredom and Baseline (0.13) and between Frustration and Boredom (0.11). This could imply that the model requires further refinement to better capture the distinctive features of each affective state. Future work could explore the incorporation of additional input features, architectural modifications, or the utilization of more sophisticated training techniques to improve classification performance.

Overall, the results suggest that the game was effective in inducing different emotions in participants, and that the combination of thermal imaging and action units with the 7-s window was the most effective modality for emotion classification. This is consistent with previous research [42], which shows that 7-s windows are optimal for thermal imaging.

In addition, the LSTM model generally performed better than the GNB model, achieving higher accuracy and F1 scores across all modalities and window sizes. Specifically, the LSTM model achieved an accuracy of 77% and an F1 score of 78% with a standard deviation of 13% when using the multimodal approach (Thermal + Action Units) and a window size of 7 s, whereas the GNB model achieved an accuracy of 35% and an F1 score of 38% with a standard deviation of 17% under the same conditions.

Our findings indicate that Long Short-Term Memory (LSTM) models outperform Gaussian Naïve Bayes (GNB) due to their ability to handle sequence data and address long-term dependency problems. The GNB's assumptions of feature independence and lack of temporal information modeling contribute to its poor performance.

To prevent overfitting, we employed cross-validation, regularization, and performance monitoring techniques. Furthermore, incorporating thermal imaging data alongside facial action units (FAUs) has proven beneficial in situations where facial expressions alone are insufficient for accurate affect detection. This additional data helps capture physiological changes, such as changes in blood flow, skin temperature, and sweating, leading to a more comprehensive understanding of affective states and improved model performance, particularly in cases with subtle expressions or challenging visual conditions.

The literature on facial action units (AUs) and their association with emotional states such as frustration, boredom, and enjoyment are inconclusive and task-dependent. Studies have reported various AUs as indicators of these emotions, such as AU02, AU09, AU14, AU17, AU18, and AU24 for frustration [61], AU04, AU26, AU07, and AU12 for boredom [62], and AU07, AU12, AU25, and AU26 for enjoyment [63]. Our results, depicted in Fig. 8, reveal that the feature AU26 (jaw drop) was the most highly ranked, indicating that it holds significant importance in the model's prediction. Given that AU26 is listed as an indicator of both boredom and enjoyment in the literature, it may suggest that this feature carries significant weight in predicting these emotions in the specific task considered in this study. However, it is important to acknowledge that the significance and relevance of the features may vary across different tasks and studies.

Additionally, the results of permutation importance analysis revealed that the minimum of the forehead region was the most important thermal feature. This feature was followed by the difference of the lower lip region, which is in alignment with previous studies [42] that have also identified the lower lip as an important feature in the facial thermal region. However, it is noteworthy that the minimum of the thermal regions was not previously identified as an important feature in the literature. Our study highlights the potential usefulness of this feature, which may be an area for further investigation in future studies.

Overall, these results provide valuable insights into the relative importance of each feature category in the model. Understanding the importance of each feature group can inform feature selection decisions in future model development efforts, as it allows us to focus on the most relevant and informative features. Additionally, these results can be utilized to identify potential areas for improvement in the model, by focusing on the features with the lowest permutation importance scores and finding ways to incorporate them more effectively into the model.

6 Limitations

A limitation of this study is that it only examined one game and one scenario, so the findings might not be generalized to other games or scenarios. Additionally, the self-report data relied on participants' subjective assessments of their own emotions, which may not always be accurate. The use of a 5-point scale may also not have provided sufficient granularity to accurately measure the differences in emotions.

The machine learning models used in this study were only trained on a small sample of participants, which may not be representative of the wider population. The accuracy of the models may also not generalize to other contexts or individuals with physical or cognitive difficulties. Furthermore, the use of thermal imaging and action units as modalities for emotion classification have also their own limitations. Such as thermal imaging is sensitive to external factors and can be affected by clothing and other objects covering the skin, while action units may not be detectable in all individuals and can be influenced by facial expressions that are not related to emotions.

A key limitation of this study is the narrow demographics of the participant sample. While gender was balanced (53% male, 47% female), factors like age, race, and ethnicity can significantly impact facial thermoregulation and the thermal expression of emotion. For example, studies show that aged skin has decreased ability to constrict facial blood vessels in response to emotional arousal, potentially reducing the magnitude of thermal signals [64]. Similarly, the density of arteriovenous anastomoses in the cheeks, which are primarily responsible for emotional blushing and temperature changes, have been shown to differ across ethnic groups [65]. These demographic variations suggest that the model performance observed here may not generalize to more diverse populations. A wider, more representative sample is needed to understand how factors like age, race, and ethnicity may influence the relationship between affective state and facial thermal patterns.

In this study, another limitation was the lack of fine-tuning the labeling of the emotions for each gameplay. If emotions were labeled after each game, this could potentially induce boredom in participants. Fine-tuning the labels per game could be problematic as it would require a significant amount of intervention, which could itself induce overall frustration during the experimental flow.

Future work in this area could include examining the level of boredom and frustration over time during gameplay, as well as studying the effects of different games and scenarios on emotions. This could provide a more comprehensive understanding of the impact of games on emotions and how to effectively measure them.

Affective states are complex, multi-dimensional experiences that can be represented in different ways. One common method is to use a two-dimensional emotion space, with valence (positive or negative) and arousal (high or low) as the axes. It is generally advisable to include valence in a two-dimensional affective state representation, as it is an important aspect of emotion and plays a role in various psychological processes and behaviors. However, there may be some cases where it is acceptable to omit valence from the representation, such as when the focus of the study is on a particular emotional experience that is not characterized by valence, such as enjoyment, frustration, or boredom. In these cases, valence may not be necessary for accurately representing the emotion.

We used a simple concatenation approach to fuse the thermal and facial action unit modalities. While this approach has been used in previous studies, it may not be the most effective way to combine multiple modalities. There are other fusion methods, such as attention-based approaches [66] and graph convolutional networks [67], that have been shown to improve the performance of multimodal affect recognition models.

Lastly, there are many state-of-the-art models that can better leverage the rich information present in thermal images and FAUs. While our study focused on demonstrating the potential of thermal imaging as a new modality for affect recognition, we recognize that future research should explore more advanced models that can better integrate multiple modalities and provide more accurate predictions. Another contribution to accurate predictions is the way self-reports are collected. We relied on self-reports at the end of the game session to determine the overall emotional state of the participants. Although this is a common approach in affective computing studies, it has limitations as it may not capture the nuances of the emotional experience during the game. Using a more fine-grained measure of emotions in real-time could have provided a more accurate understanding of affective states during the game.

7 Conclusion

In conclusion, this study aimed to explore the effectiveness of using thermal imaging, facial action units, and a combination of both, for the detection of four distinct emotional states (frustration, boredom, enjoyment, and neutral) during a tangible gamified exercise. The self-reports showed that the game was successful in inducing significant differences between the phases.

The machine learning models that combined thermal imaging and action units data achieved the highest accuracy of 77% in affect classification within a 7-s window and while using only thermal data had lower standard deviation among participants.

In the sphere of social robotics, the ability to accurately detect and respond to human emotions is crucial for fostering effective human–robot interactions. By integrating the proposed multimodal approach into social robots, these intelligent systems can better understand and adapt to the affective states of their users, thereby enhancing the overall therapeutic experience. For instance, social robots can modify their behavior and therapeutic strategies based on the detected emotions, leading to more personalized and engaging therapy sessions.

Furthermore, the results of permutation importance analysis showed significant features for affect classification like AU26 and minimum forehead temperature and correlations to the current literature. Overall, the present study highlights the need for further research to explore the potential utility of thermal imaging and action units in designing affect-aware healthcare technologies with the target groups.

Acknowledgements This work was partially funded by grants from the Digital Futures Research Center. This research received additional funding from the H2020 ACCIO TecnioSpring INDUSTRY project (grant agreement no. 801342). I, Youssef Mohamed, hereby grant permission to the International Journal of Social Robotics to publish the manuscript mentioned above in print and electronic format. I affirm that the manuscript is original and has not been published elsewhere. I also affirm that all the authors listed in the manuscript have contributed significantly to the work and have agreed to its submission for publication.

Funding Open access funding provided by Royal Institute of Technology.

Declarations

Ethical approval According to the national regulations in the country where this research was conducted, we are exempt from ethical approval because we did not collect any sensitive personal data (racial or ethnic origin, political views, religious or philosophical beliefs, health or sexual life), and our research does not involve physical intervention on the research person or biological samples from participants. An informed consent form is prepared according to the KTH's ethical regulations and each participant signed this informed consent form before the experiment. The authors declare that they have no conflict of interest. Anonymized and unidentifiable data could be available on request (action units and thermal facial region values) within the research community.

References

1. Leite D, Frigeri V, Medeiros R (2021) Adaptive gaussian fuzzy classifier for real-time emotion recognition in computer games. In: 2021 IEEE Latin American conference on computational intelligence (LA-CCI), pp 1–6. https://doi.org/10.1109/LA-CCI48322.2021.9769842
2. Tijs T, Brokken D, IJsselsteijn W (2009) Creating an emotionally adaptive game. In: Stevens SM, Saldamarco SJ (eds) Entertainment Computing—ICEC 2008. Springer, Berlin, pp 122–133
3. Tivatansakul S, Ohkura M, Puangpontip S, Achalakul T (2014) Emotional healthcare system: emotion detection by facial expressions using Japanese database. In: 2014 6th computer science and electronic engineering conference (CEEC). IEEE, pp 41–46
4. Szwoch M, Szwoch W (2015) Emotion recognition for affect aware video games. In: Image processing and communications challenges 6. Springer, pp 227–236
5. Csikszentmihalyi M, Csikzentmihaly M (1990) Flow: the psychology of optimal experience, vol. 1990. Harper & Row, New York

6. Vaughan N, Gabrys B, Dubey VN (2016) An overview of self-adaptive technologies within virtual reality training. Comput Sci Rev 22:65–87
7. Novais P, Carneiro D (2016) The role of non-intrusive approaches in the development of people-aware systems. Prog Artif Intell 5(3):215–220
8. Tian L, Oviatt S, Muszynski M, Chamberlain BC, Healey J, Sano A (2022) Emotion-aware human–robot interaction and social robots. Appl Affect Comput
9. McDuff D, Czerwinski M (2018) Designing emotionally sentient agents. Commun ACM 61(12):74–83
10. Gilleade KM, Dix A (2004) Using frustration in the design of adaptive videogames. In: Proceedings of the 2004 ACM SIGCHI international conference on advances in computer entertainment technology. ACE '04. Association for Computing Machinery, New York, NY, USA, pp 228–232. https://doi.org/10.1145/1067343.1067372
11. Park S, Lee SW, Whang M (2021) The analysis of emotion authenticity based on facial micromovements. Sensors 21(13):4616
12. Nomura K, Iwata M, Augereau O, Kise K (2019) Estimation of student's engagement based on the posture. In: Adjunct proceedings of the 2019 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2019 ACM international symposium on wearable computers, pp 164–167
13. Zhao S, Wang S, Soleymani M, Joshi D, Ji Q (2019) Affective computing for large-scale heterogeneous multimedia data: a survey. Assoc Comput Mach. https://doi.org/10.1145/3363560
14. Picard RW (2003) Affective computing: challenges. Int J Hum Comput Stud 59(1–2):55–64. https://doi.org/10.1016/S1071-5819(03)00052-1
15. Kothig A, Munoz J, Akgun SA, Aroyo AM, Dautenhahn K (2021) Connecting humans and robots using physiological signals—closing-the-loop in HRI. pp 735–742. https://doi.org/10.1109/ro-man50785.2021.9515383. https://www.researchgate.net/publication/354081910
16. Cross CB, Skipper JA, Petkie D (2013) Thermal imaging to detect physiological indicators of stress in humans. In: Thermosense: thermal infrared applications XXXV, vol. 8705. SPIE, p 87050. https://doi.org/10.1117/12.2018107. https://www.spiedigitallibrary.org/terms-of-use
17. Stemberger J, Allison RS, Schnell T (2010) Thermal imaging as a way to classify cognitive workload. In: CRV 2010—7th Canadian conference on computer and robot vision, pp 231–238. https://doi.org/10.1109/CRV.2010.37
18. Abdelrahman Y, Velloso E, Dingler T, Schmidt A, Vetere F (2017) Cognitive heat. Proc ACM Interact Mob Wearable Ubiquitous Technol 1(3):1–20. https://doi.org/10.1145/3130898
19. Shastri D, Merla A, Tsiamyrtzis P, Pavlidis I (2009) Imaging facial signs of neurophysiological responses. IEEE Trans Biomed Eng 56(2):477–484. https://doi.org/10.1109/TBME.2008.2003265
20. Sinha R, Lovallo WR, Parsons OA (1992) Cardiovascular differentiation of emotions. Psychosom Med 54(4):422–435. https://doi.org/10.1097/00006842-199207000-00005
21. Collet C, Vernet-Maury E, Delhomme G, Dittmar A (1997) Autonomic nervous system response patterns specificity to basic emotions. J Auton Nerv Syst 62(1–2):45–57. https://doi.org/10.1016/S0165-1838(96)00108-7
22. Guneysu Ozgur A, Wessel MJ, Johal W, Sharma K, Özgür A, Vuadens P, Mondada F, Hummel FC, Dillenbourg P (2018) Iterative design of an upper limb rehabilitation game with tangible robots. In: Proceedings of the 2018 ACM/IEEE international conference on human–robot interaction, pp 241–250
23. Weidemann A, Rußwinkel N (2021) The role of frustration in human-robot interaction—What is needed for a successful collaboration? Front Psychol 12:707. https://doi.org/10.3389/fpsyg.2021.640186
24. Kapoor A, Burleson W, Picard RW (2007) Automatic prediction of frustration. Int J Hum Comput Stud 65(8):724–736. https://doi.org/10.1016/j.ijhcs.2007.02.003
25. Taylor B, Dey A, Siewiorek D, Smailagic A (2015) Using physiological sensors to detect levels of user frustration induced by system delays. In: UbiComp 2015—Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing. Association for Computing Machinery, Inc, pp 517–528. https://doi.org/10.1145/2750858.2805847
26. Bosch N, Chen H, D'Mello S, Baker R, Shute V (2015) Accuracy vs. availability heuristic in multimodal affect detection in the wild. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp 267–274
27. Shibasaki Y, Funakoshi K, Shinoda K (2017) Boredom recognition based on users' spontaneous behaviors in multiparty human–robot interactions. In: International conference on multimedia modeling. Springer, pp 677–689
28. Lewinski P, Den Uyl TM, Butler C (2014) Automated facial coding: validation of basic emotions and FACS AUs in FaceReader. J Neurosci Psychol Econ 7(4):227
29. De Silva LC, Miyasato T, Nakatsu R (1997) Facial emotion recognition using multi-modal information. In: Proceedings of ICICS, 1997 international conference on information, communications and signal processing. Theme: trends in information systems engineering and wireless multimedia communications (Cat.), vol 1. IEEE, pp 397–401
30. Wang Z, Ho S-B, Cambria E (2020) A review of emotion sensing: categorization models and algorithms. Multimed Tools Appl 79(47):35553–35582
31. Alonso-Martin F, Malfaz M, Sequeira J, Gorostiza JF, Salichs MA (2013) A multimodal emotion detection system during human–robot interaction. Sensors 13(11):15549–15581
32. Psaltis A, Kaza K, Stefanidis K, Thermos S, Apostolakis KC, Dimitropoulos K, Daras P (2016) Multimodal affective state recognition in serious games applications. In: 2016 IEEE international conference on imaging systems and techniques (IST), pp 435–439. https://doi.org/10.1109/IST.2016.7738265
33. Fydanaki A, Geradts Z (2018) Evaluating OpenFace: an open-source automatic facial comparison algorithm for forensics. Forensic Sci Res 3(3):202–209. https://doi.org/10.1080/20961790.2018.1523703
34. Lloyd JM (1975) Thermal imaging systems. Springer, Boston, pp 1–17. https://doi.org/10.1007/978-1-4899-1182-7_1
35. Nguyen T, Tran K, Nguyen H (2018) Towards thermal region of interest for human emotion estimation. In: Proceedings of 2018 10th international conference on knowledge and systems engineering, KSE 2018, pp 152–157. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/KSE.2018.8573373
36. Ioannou S, Gallese V, Merla A (2014) Thermal infrared imaging in psychophysiology: potentialities and limits. Psychophysiology 51(10):951–963. https://doi.org/10.1111/psyp.12243
37. Cho Y, Bianchi-Berthouze N, Oliveira M, Holloway C, Julier S (2019) Nose heat: exploring stress-induced nasal thermal variability through mobile thermal imaging. In: 2019 8th international conference on affective computing and intelligent interaction (ACII). IEEE, pp 566–572
38. Cho Y, Julier SJ, Bianchi-Berthouze N (2019) Instant stress: detection of perceived mental stress through smartphone photoplethysmography and thermal imaging. JMIR Ment Health 6(4):10140. https://doi.org/10.2196/10140
39. Engert V, Merla A, Grant JA, Cardone D, Tusche A, Singer T (2014) Exploring the use of thermal infrared imaging in human stress research. PLOS ONE 9(3):90782
40. Veltman HJ, Vos WW (2005) Facial temperature as a measure of mental workload. In: 2005 International symposium on aviation psychology, p 777

41. Sorostinean M, Ferland F, Tapus A (2015) Reliable stress measurement using face temperature variation with a thermal camera in human-robot interaction. In: IEEE-RAS international conference on humanoid robots, vol 2015-December. IEEE Computer Society, pp 14–19. https://doi.org/10.1109/HUMANOIDS.2015.7363516
42. Mohamed Y, Ballardini G, Parreira MT, Lemaignan S, Leite I (2022) Automatic frustration detection using thermal imaging. In: Proceedings of the 2022 ACM/IEEE international conference on human–robot interaction, pp 451–460
43. Kort B, Reilly R, Picard RW (2001) An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion. In: Proceedings IEEE international conference on advanced learning technologies. IEEE, pp 43–46
44. Özgür A, Lemaignan S, Johal W, Beltran M, Briod M, Pereyre L, Mondada F, Dillenbourg P (2017) Cellulo: versatile handheld robots for education. In: 2017 12th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 119–127
45. Olsen JK, Guneysu Ozgur A, Sharma K, Johal W (2022) Leveraging eye tracking to understand children's attention during game-based, tangible robotics activities. Int J Child Comput Interact 31:100447
46. Guneysu Ozgur A, Wessel MJ, Olsen JK, Cadic-Melchior AG, Zufferey V, Johal W, Dominijanni G, Turlan J-L, Mühl A, Bruno B (2022) The effect of gamified robot-enhanced training on motor performance in chronic stroke survivors. Heliyon 8(11):11764
47. Guneysu Ozgur A, Wessel MJ, Olsen JK, Johal W, Ozgur A, Hummel FC, Dillenbourg P (2020) Gamified motor training with tangible robots in older adults: a feasibility study and comparison with the young. Front Aging Neurosci 12:59
48. Guneysu Ozgur A, Wessel MJ, Asselborn T, Olsen JK, Johal W, Özgür A, Hummel FC, Dillenbourg P (2019) Designing configurable arm rehabilitation games: How do different game elements affect user motion trajectories? In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 5326–5330
49. Dollard J, Miller NE, Doob LW, Mowrer OH, Sears RR (1939) Frustration and aggression. Yale University Press. https://doi.org/10.1037/10022-000
50. Hart SG, Staveland LE (1988) Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv Psychol 52(C):139–183. https://doi.org/10.1016/S0166-4115(08)62386-9
51. Markland D, Hardy L (1997) On the factorial and construct validity of the intrinsic motivation inventory: conceptual and operational concerns. Res Q Exerc Sport 68(1):20–32
52. Ekman P (2003) Emotions revealed: recognizing faces and feelings to improve communication and emotional life, p 285. https://psycnet.apa.org/record/2003-88051-000
53. Stiber M, Taylor R, Huang C-M (2022) Modeling human response to robot errors for timely error detection. arXiv:2208.00565
54. Do N-T, Nguyen-Quynh T-T, Kim S-H (2020) Affective expression analysis in-the-wild using multi-task temporal statistical deep learning model. In: 2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020), pp 624–628. https://doi.org/10.1109/FG47880.2020.00093
55. Hung JC, Xu S-L (2022) Analysis for sequential frame with facial emotion recognition based on CNN and LSTM. In: International conference on innovative computing. Springer, pp 112–122
56. Yücetürk NE, Demir S, Özdemir Z, Bejan I, Drešević N, Katanić M, Dillenbourg P, Soysal A, Ozgur AG (2022) Predictive analysis of errors during robot-mediated gamified training. In: 2022 International conference on rehabilitation robotics (ICORR). IEEE, pp 1–6
57. Jahromi AH, Taheri M (2017) A non-parametric mixture of Gaussian Naive Bayes classifiers based on local independent features. In: 2017 Artificial intelligence and signal processing conference (AISP). IEEE, pp 209–212
58. Cohen I, Sebe N, Garg A, Chen LS, Huang TS (2003) Facial expression recognition from video sequences: temporal and static modeling. Comput Vis Image Underst 91(1–2):160–187
59. Sebe N, Lew MS, Cohen I, Garg A, Huang TS (2002) Emotion recognition using a Cauchy Naive Bayes classifier. In: Object recognition supported by user interaction for service robots, vol 1. IEEE, pp 17–20
60. Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20(177):1–81
61. Ihme K, Unni A, Zhang M, Rieger JW, Jipp M (2018) Recognizing frustration of drivers from face video recordings and brain activation measurements with functional near-infrared spectroscopy. Front Hum Neurosci 12:327
62. D'Mello S, Craig S, Gholson B, Franklin S, Picard R, Graesser A (2004) Integrating affect sensors into an intelligent tutoring system. In: Affective interactions: the computer in the affective loop. Proceedings of the 2005 international conference on intelligent user interfaces, pp 7–13
63. McDaniel B, D'Mello S, King B, Chipman P, Tapp K, Graesser A (2007) Facial features for affective state detection in learning environments. In: Proceedings of the annual meeting of the cognitive science society, vol 29
64. Frank SM, Raja SN, Bulcao C, Goldstein DS (2000) Age-related thermoregulatory differences during core cooling in humans. Am J Physiol Regul Integr Comp Physiol 279(1):349–354
65. Drummond PD, Lim HK (2000) The significance of blushing for fair-and dark-skinned people. Personal Individ Differ 29(6):1123–1132
66. Huddar MG, Sannakki SS, Rajpurohit VS (2021) Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN
67. Duhme M, Memmesheimer R, Paulus D (2022) Fusion-GCN: multimodal action recognition using graph convolutional networks. In: Pattern recognition: 43rd DAGM German conference, DAGM GCPR 2021, Bonn, Germany, September 28–October 1, 2021, Proceedings. Springer, pp 265–281

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Youssef Mohamed I am a PhD student at the Division of Robotics, Perception and Learning, KTH Royal Institute of Technology. I have done my BSc in Mechanical engineering at the University of Surrey and focused mainly on autonomous vehicles and fuzzy controllers. Then did my MSc in Robotics at the Univesity of Bristol, with a focus on social robotics and human-robotinteraction.

Arzu Güneysu I am a lecturer at the Department of Biomedical Engineering and Health Systems. I was a postdoctoral researcher at Digital Futures in the Division of Robotics Perception and Learning at KTH. Igot my Ph.D. in Robotics on "Designing Gamified Activities with Haptic-Enabled Tangible Robots for Therapyand Assistance" from EPFL, in 2021. My research interests include various topics in Human-Robot Interaction, Adaptive Robot-Enhanced Therapy, Iterative Design, Participatory Design, Neuro developmental Disorders, Gamified Therapeutic Technologies, Healthy Aging, Intergenerational Practices for the Elderly and Children, and Special Education. My supervisors were Iolanda Leite, Associate Professor, at the Department of Robotics,Perception and Learning at KTH, and Ali Reza Majlesi, Associate Professor, at the Department of Education at Stockholm University.

Severin Lemaignan Since 2021, I’m Head of Social Robotics and AI and Senior Scientist at Barcelona-based PAL Robotics, one of the EU leader in service robotics. I lead the Human-Robot Interaction group. Previously, I was Associate Prof in Social Robotics and AI at the Bristol Robotics Lab, UK. My expertise focuses on cognitive and social robotics, with a particular focus on socially assistive robotics.

Iolanda leite I am an Associate Professor at the Division of Robotics, Perception and Learning. I received my PhD degree from the Technical University of Lisbon. Before joining KTH, I was a Postdoctoral Associate atthe Yale Social Robotics Lab and an Associate Research Scientist at Disney Research. The goal of my researchis to develop social robots that cancapture, learn from and respond appropriately to the subtle dynamics that characterize real-world situations, allowing for truly efficient and engaging long-term interactions with people.