IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 July 2023, accepted 24 August 2023, date of publication 1 September 2023, date of current version 7 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3311268

RESEARCH ARTICLE

Multimodal Assessment of Interest Levels in Reading: Integrating Eye-Tracking and Physiological Sensing

JAYASANKAR SANTHOSH[1,2], DAVID DZSOTJAN[1], AND SHOYA ISHIMARU[1,2], (Member, IEEE)
[1]Department of Computer Science, RPTU Kaiserslautern-Landau, 67663 Kaiserslautern, Germany
[2]German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany

Corresponding author: Jayasankar Santhosh (jayasankar.santhosh@dfki.de)

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG) International Call on Artificial Intelligence ‘Learning Cyclotron’.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by DFKI Ethics Board.

ABSTRACT Interest in the context of reading holds special significance as it serves as a driving force for learning and education. By understanding and leveraging students’ interests, educators can create more effective and enjoyable learning environments that promote personalized learning experiences, enhanced comprehension, deep understanding, and motivation. A multimodal approach integrating gaze and physiological data could provide a more comprehensive and accurate assessment of interest levels. The goal of this study is to measure the level of interest experienced by users when reading newspaper articles by integrating gaze data and physiological responses. An experiment was conducted which recorded the gaze and physiological data from 13 university students reading 18 newspaper articles collected from the BBC news database. An SMI eye-tracker and an Empatica E4 wristband were used synchronously to capture the user’s eye movements and physiological data. To predict the interest levels of the participants, a manual feature extraction-based approach and a deep learning-based approach were employed. The interest levels were divided into four-class and binary based on the responses from the participants. A CNN-LSTM model using the gaze features outperformed other models in terms of accuracy and F1-score with 52.8% and 51.8 for four-class and 82.3% and 81.7 for binary classification using leave-one-document-out and leave-one-participant-out cross-validation, respectively.

INDEX TERMS Affective-computing, eye-tracking, physiological sensing, reading analysis.

I. INTRODUCTION

Understanding what it is that drives humans, how motivation works, and what interest really is, has been a subject of research since well before the rise of digital technology. Interest, a word so often used, is nonetheless difficult to express as a single concept. Therefore several definitions exist based on aspect and context. In the fields of psychology, sociology, and education research, it is generally accepted that interest is “a construct that characterizes a person’s special relationship with an object (contents, topics, special subject, and object domain)” [1]. Numerous researchers link it to academic achievement [2], [3], [4], or define it as a motivational variable that is the foundation of productive learning [5], [6], [7].

Interest is also studied in terms of recall: information that an individual finds interesting is longer and more easily retained than non-interesting information [8]. According to Hockenbury and Hockenbury, emotions consist of the following three components: physiological response, expressive behavioral response, and subjective experience [9]. As the results in the works of Izard et al., Silvia et al., and Libby et al. suggest, interest seems to encompass these qualities as well [5], [10], [11], [12]. Thus, physiological outputs of the

The associate editor coordinating the review of this manuscript and approving it for publication was Angelo Trotta.

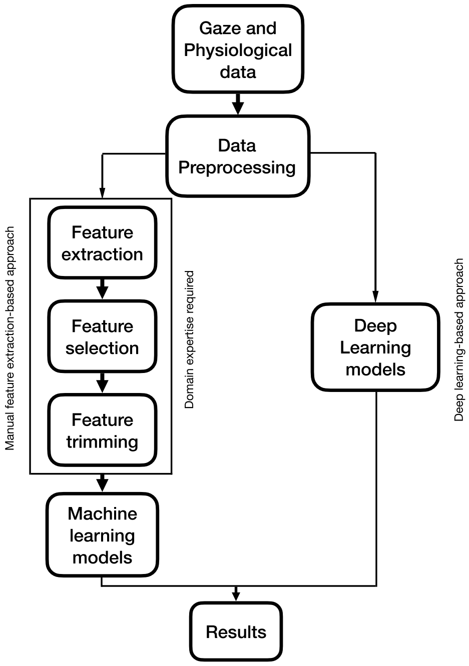

Gaze and Physiological data
Data Preprocessing
Feature extraction
Feature selection
Feature trimming
Machine learning models
Deep Learning models
Results
Manual feature extraction-based approach
Domain expertise required
Deep learning-based approach

FIGURE 1. **An overview of manual feature extraction-based approach and deep learning-based approach.**

human body provide enticing access to measure interest and motivation.

Interest in the context of reading has special importance, since learning and education are predominantly accomplished through the act of reading. Understanding what facilitates the urge to read can greatly help in designing documents that make for a more efficient human-document interaction. The cognitive processes in the mind of a reader are mirrored by their eyes: prompting changes in pupil diameter, blinks, fixations, saccades, and regressions [13], [14]. Thus, eye-tracking is a straightforward choice for interest detection in reading tasks. Heart rate behavior is another physiological channel where emotions and thus interest, too, can be efficiently measured [15], [16], [17], [18]. Using a wristband as a detector has the advantage of being extremely unobtrusive, and thus introduces practically no distraction during the reading task. The findings of the research by Giradi et al. demonstrated that the emotions of developers while programming can be detected using a minimal set of biometric features and sensors put on a single wearable device (Empatica E4) [19]. The minimum set of sensors measured using the Empatica E4 wristband was successfully identified, which can be used in an experimental protocol for detecting emotions during software development tasks [19].

Autonomous nervous system impulses are consistently the building blocks of cognition or emotion. Heart rate variability, electrodermal activity, and skin temperature are physiological indicators of an individual's affective state during moments of arousal, engagement/interest, boredom, and anger. Research has been conducted on the relationship between various emotional states, engagement, and learning to empower the experience of learning. These data have previously been measured via Electroencephalogram (EEG), Electrocardiogram (ECG), Electrooculography (EOG), Galvanic Skin Response (GSR), Eye-Tracking (ET) and Electrodermal activity (EDA) but wristbands today collect physiological data equally effectively. Such portable and unobtrusive devices allow data collection with reduced restraints compared to traditional laboratory devices. Brishtel et al. evaluated the disparities in information processing on screen and paper using an E4 wristband and an eye tracker, and the results indicated that portable devices such as the E4 are well tailored for monitoring mental workload [20].

In our previous two studies, data from the eye tracker and Empatica E4 wristband were separately evaluated using conventional machine learning methods by manually extracting features from gaze data [21], and physiological signals [22]. The fundamental objective of this study is to have different degrees of interest identified using the raw attributes of the eye tracker and the Empatica E4 wristband for the same recorded data. Additionally, the previously achieved outcomes using conventional machine learning-based methods will be compared to the outcomes obtained through the deep-learning based approach. The deep learning models, such as a One-dimensional Convolutional neural network (1D CNN), CNN-LSTM (long short-term memory) [23], and a Fully Convolutional Network (FCN) [24] were implemented to extract high-level features from raw sensor input, minimizing the strenuous task of explicitly extracting features from the gaze and physiological data.

In the prior work, interest classification was computed using conventional machine learning models such as Random Forest (RF) and Support Vector Machine (SVM), which required manual feature extraction and domain expertise. Meanwhile, deep learning models have the benefit of extracting high-level features from raw sensory input and generating better outcomes [25], [26], [27], [28], [29]. The proposed method can handle temporal dependencies in the data, which is important for physiological data classification tasks and can learn from data with different levels of abstraction, allowing them to capture both low-level and high-level features in the data. Figure 1 provides an overview or flowchart of the proposed system and facilitates a comparison between two distinct approaches: the conventional machine learning-based approach and the deep learning-based approach for interest detection. It illustrates the flow of data and processing steps involved in the proposed system, emphasizing the main stages and their connections. Many of the prior studies for cognitive/affective state detection utilize traditional manual feature extraction-based approach, and the effectiveness of using deep learning models are less explored. Therefore, the following research aims are the main focus of this work.

- Predicting reader interest using a deep learning-based approach and comparing the outcomes with a manual feature extraction-based approach.
- Evaluating whether combining multiple sensing modalities could leverage cognitive state prediction in a person.

- Examining how person-specific predictions compare against generalized prediction models.

In order, the paper is structured as follows: Section II offers an explanation of the technical background and prior research in the domain of detecting motivation and interest through sensor-based approaches. Section III details the techniques used to examine the interest in reading, including eye-tracking and physiological data analysis. Section IV summarizes the assessment methods used and the results obtained. Section V explores the study's research inquiries, challenges, and constraints. Lastly, Section VI concludes the paper.

II. RELATED WORK

The physiological aspect of interest offers an attractive possibility for measurement using biological sensors - and indeed, there are quite a few studies where sensor data has been used to gauge interest and motivation in humans [39].

Since the human body has numerous physiological outputs where the interest can potentially be measured, it seems logical to assume that the combination of sensor data from these different channels could enhance the efficiency of interest detection. Whether that is always so, is not a straightforward task to answer, given the potentially differing time scales and varying noisiness of these modalities. That is why an important part of our work consists of examining the classification accuracy in the case of excluding individual modalities from the physiological signals.

Numerous studies aiming for the detection of interest, motivation, or cognitive states in general have been found, where a fusion of outputs from different physiological sensors is used (Table 1). These studies investigate the respective emotions not only through the use of different techniques but also in different contexts. Since the various settings emphasize different cognitive functions, and thus interest, motivation, and related emotions potentially manifest differently, each of these studies constitutes another piece in the greater puzzle.

In some of these settings, participants are typically less active and have more of the role of an observer. For instance, Suzuki et al. and Shigemitsu et al. both measure interest levels for participants watching movies using EEG and EOG signals [30], [31]. Vecchiato et al. examine the effect of interruptions in visual materials: here, participants watch documentaries that are interrupted by commercials, and the gender-based differences in the resulting positive/negative emotions, interest, and memory are detected [33].

Other studies measure interest and motivation in settings where a more active, more creative disposition is encouraged in the participants, such as reading and learning. Azcarraga et al. measure interest, excitement, frustration, as well as confidence when participants are figuring out mathematical problems, and academic emotions are predicted based on brainwaves and mouse activity [32]. Asteriadis et al. estimate user behavior in an e-learning environment, based on eye gaze and head pose detection, by measuring whether the user is frustrated or struggling to read, is distracted, tired/sleepy, is not paying attention, or is full of interest [34]. By both Mota et al. and Kapoor et al., the learners' posture analysis (based on video monitoring and pressure sensors in the chair) is used to gauge the level of interest while solving a constraint satisfaction game [35], [36]. However, while the former's primary interest lies in the detection of different interest levels (high, medium, or low interest, taking a break, and boredom), the latter's main concern is the feasibility of using a combination of different modalities to detect interest/disinterest. In addition, Mota et al. also use computer screen activity data for the detection, while by Kapoor et al. the game state information is a non-physiological modality to detect interest.

The goal of Arroyo et al. is not only the detection of the emotional state (interest, frustration, excitement, confidence) of the learner while using an adaptive multimedia tutoring system but also the application of the measured information by providing adaptive emotional support to the student [37]. The focus of both Ishimaru et al. and Brishtel et al. is the detection of attention and cognitive states in a learning scenario in the context of human-document interaction [20], [38]. In particular, in the former study, the cognitive states of learners are quantified based on the signals from an eye tracker and a thermal camera while they are reading a textbook, while the authors in the latter investigate the effects of text semantics and music on attention and mind wandering while reading a document.

Researchers have utilized eye tracking technology and physiological sensors to gather data and infer readers' cognitive and affective states. Eye tracking technology has been employed to monitor eye movements and gaze patterns during reading. By analyzing fixations, saccades, and reading speed, researchers have attempted to identify markers of attention, comprehension, and interest. Changes in eye movement patterns, such as increased fixations or longer dwell times on specific text passages, can indicate heightened interest or cognitive processing.

Physiological sensing techniques, including measures like heart rate, skin conductance, and facial expressions have also been employed to capture readers' emotional and cognitive responses during reading. These physiological signals can provide insights into readers' arousal, valence, and emotional engagement. For instance, an increased heart rate or enhanced skin conductance may indicate heightened interest or emotional response to specific text content.The combination of eye tracking and physiological sensing allows for a more comprehensive understanding of readers' interest levels. Integrating these modalities enables researchers to investigate the dynamic interplay between cognitive processing, visual attention, and emotional responses during reading. Table 2 outlines the advantages and disadvantages of employing different sensing techniques to understand users' cognitive states.

The context for interest detection in the present work is human-document interaction. Gauging the learners' interest reliably in this setting has singular importance, since the vast majority of learning at school and university happens

TABLE 1. Studies from the period 2005-2020 where a combination of different sensors was used to measure interest, motivation, and cognitive states. 'Y' (yes) and 'N' in each row represents which sensors were used and not used in these studies.

Study	EEG	EOG	ECG	GSR	ET	EDA	Video capture of			Blue eyes	Thermal	Pressure	Mouse
							Face	Hand	Posture	video	camera	sensors	activity
Suzuki et al. [30]	Y	Y	N	N	N	N	N	N	N	N	N	N	N
Shigemitsu et al. [31]	Y	Y	N	N	N	N	N	N	N	N	N	N	N
Azcarraga et al. [32]	Y	N	N	N	N	N	N	N	N	N	N	N	Y
Vecchiato et al. [33]	Y	N	Y	Y	N	N	N	N	N	N	N	N	N
Asteriadis et al. [34]	N	N	N	N	Y	N	N	Y	Y	N	N	N	N
Mota et al. [35]	N	N	N	N	N	N	Y	N	Y	Y	N	Y	N
Kapoor and Picard [36]	N	N	N	N	N	N	Y	N	N	N	N	Y	N
Arroyo et al. [37]	N	N	N	N	N	Y	Y	N	N	N	N	Y	N
Ishimaru et al. [38]	N	N	N	N	Y	N	N	N	N	N	Y	N	N
Brishtel et al. [20]	N	N	N	N	Y	Y	N	N	N	N	N	N	N

TABLE 2. The table showing the benefits and limitations of estimating users' cognitive states using various sensors.

Benefits	Limitations
Provides objective and real-time data	Technical challenges in data collection and interpretation
Offers insights into cognitive and affective states	Potential discomfort or distraction caused by the sensors
Helps identify areas of high interest or comprehension difficulties	Variability in individual responses and interpretations
Offers opportunities for personalized reading experiences	Ethical considerations regarding privacy and data usage
Can aid in the development of adaptive reading systems	Limitations in generalizing findings to real-world reading contexts

through reading. Efficient detection of interest opens the door not only for improving adaptive learning assistance in an intelligent learning system, but also for designing documents that facilitate learning performance in students.

III. METHODOLOGY

The system was designed to record eye movements and the physiological signals from the participants. An SMI REDn Scientific 60Hz remote eye tracker, along with an Empatica E4 wristband, was used to record the data. The gaze data recorded from the eye-tracker has a timestamp, left and right gaze coordinates (x and y with the screen edges as the coordinate system), and the pupil diameter of the left and right eye. The physiological data collected from the E4 wristband sensor includes 3-axis Acceleration (ACC; 32Hz), Blood Volume Pulse (BVP; 64Hz), Electrodermal activity (EDA; 4Hz), Skin Temperature (TEMP; 4Hz), and Heart Rate (HR; 1Hz). Figure 2 represents the raw sensor signals recorded from a single participant based on the ratings provided for interest for different documents. Interestingly, the EDA could be seen rising for *high-interest* response whereas decreasing for *low-interest* response, but vice versa could be observed for temperature. For the analysis, the preprocessed data was segmented using a sliding window algorithm with a window size of 10 seconds and 50% overlap. The flowchart depicting the proposed system can be seen in Figure 1.

A. MANUAL FEATURE EXTRACTION-BASED INTEREST DETECTION

Two separate feature calculations proposed by Jacob et al. [21], [22] are followed, and the concatenation of the future dimensions was performed to investigate the effect of the sensor fusion.

1) GAZE FEATURES

The raw gaze data was preprocessed to extract the fixations and saccades. From the extracted fixations and saccades, 17 features including the fixation duration, saccade length, saccade speed, regression length, regression speed, count of saccades, and regressions were computed [21]. The pupil diameter and the blink frequency were also extracted from the raw gaze data. The statistical features such as mean and standard deviation were computed for the extracted gaze features and used to train the model for interest classification. The previous works [21], [22] contain the mathematical analysis relevant to this study.

2) WRISTBAND FEATURES

The phasic and tonic components of the EDA signal were computed after preprocessing the EDA signal from the E4 wristband, as shown in Figure 3. In addition, the slope of the tonic component and the peak amplitudes for the phasic component were calculated.

From the BVP signals, the heart rate (HR) and the inter-beat intervals (IBI) were retrieved. The mean and standard deviation were computed for the BVP and HR data, and the difference in the mean and standard deviation of the HR amplitude during the task and baseline was estimated. Figure 4 shows the raw and cleaned BVP signal and the extracted heart rate data. Similarly, the mean and standard deviation of the skin temperature (TEMP) were calculated, and the slope of the TEMP signal was obtained. In total,

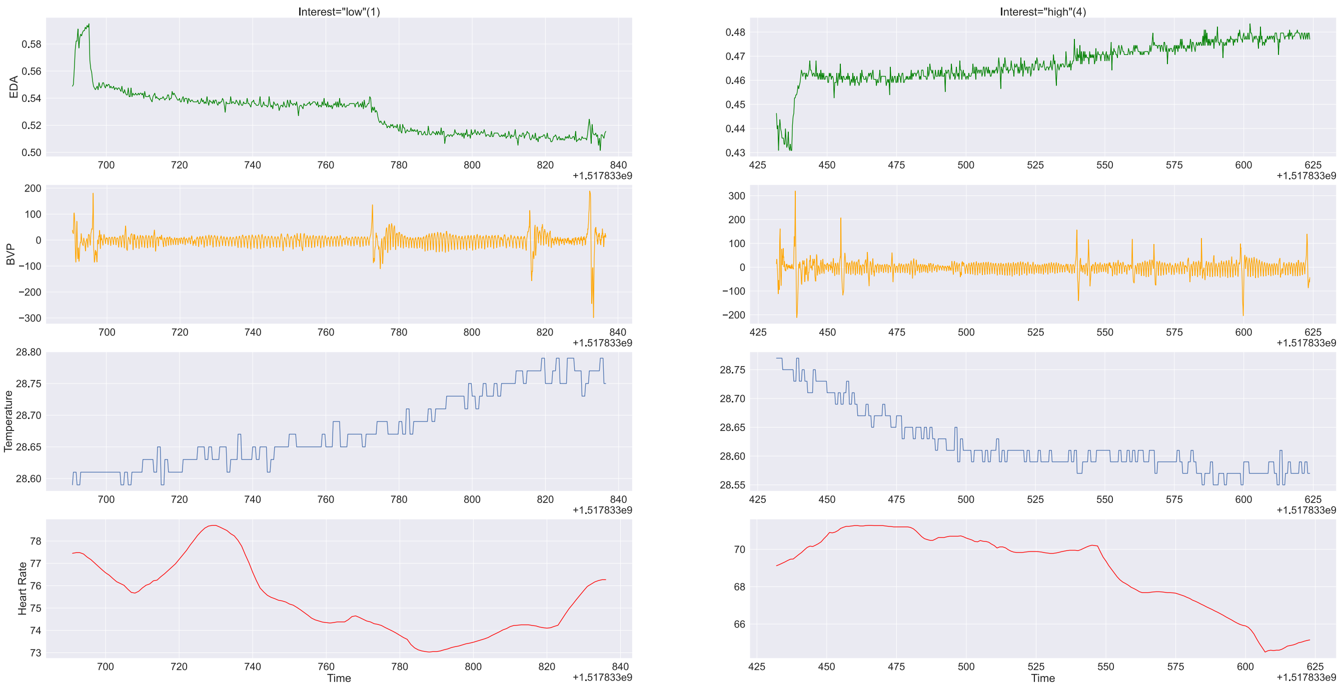


FIGURE 2. The raw physiological signals recorded from a single participant based on the interest responses for different documents.

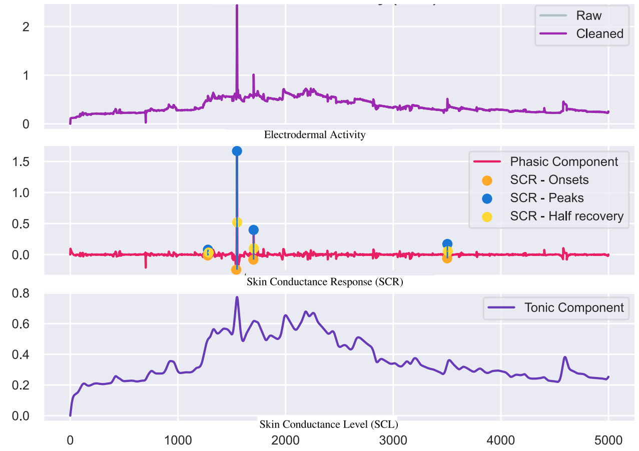


FIGURE 3. The phasic and tonic components extracted from the EDA signal.

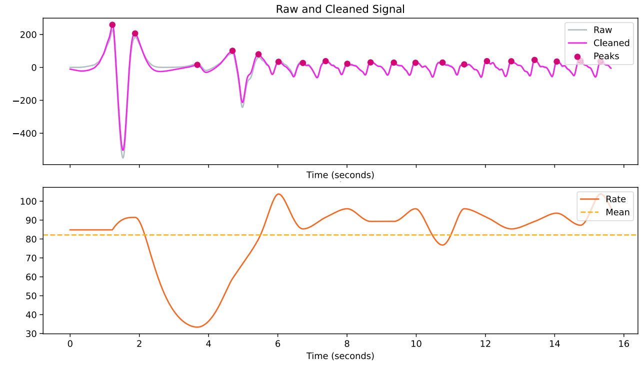


FIGURE 4. The peaks computed from the BVP signal and the extracted heart rate data.

18 statistical features listed in the original work were calculated for the classification task [22].

3) CLASSIFICATION

A Support Vector Machine (SVM) [40], [41] and a Random Forest classifier [42], [43] were utilized to predict the interest level of the participants. SVM works by finding the optimal hyperplane that maximizes the margin between the different classes of data, and then using this hyperplane to classify new data points. SVM is particularly effective in handling high-dimensional data and can handle complex data sets with a high degree of accuracy. The features from the eye tracker and features from the E4 wristband were used separately and combined as well for training the classifiers. The random forest classifier chooses random subsets of training data to build decision tree sets, and then utilizes the combined votes of these trees to determine the category of the test data.

Choosing the right hyperparameters (such as C, kernel, and gamma in SVM, and the number of estimators, maximum tree depth, maximum features, minimum samples required to split an internal node, and minimum samples required to be at a leaf node in Random Forest) is significant, but it's not an easy task. In our method, a 3-fold grid search cross-validation was employed to optimize the hyperparameters. This method exhaustively searches through a predefined set of parameters

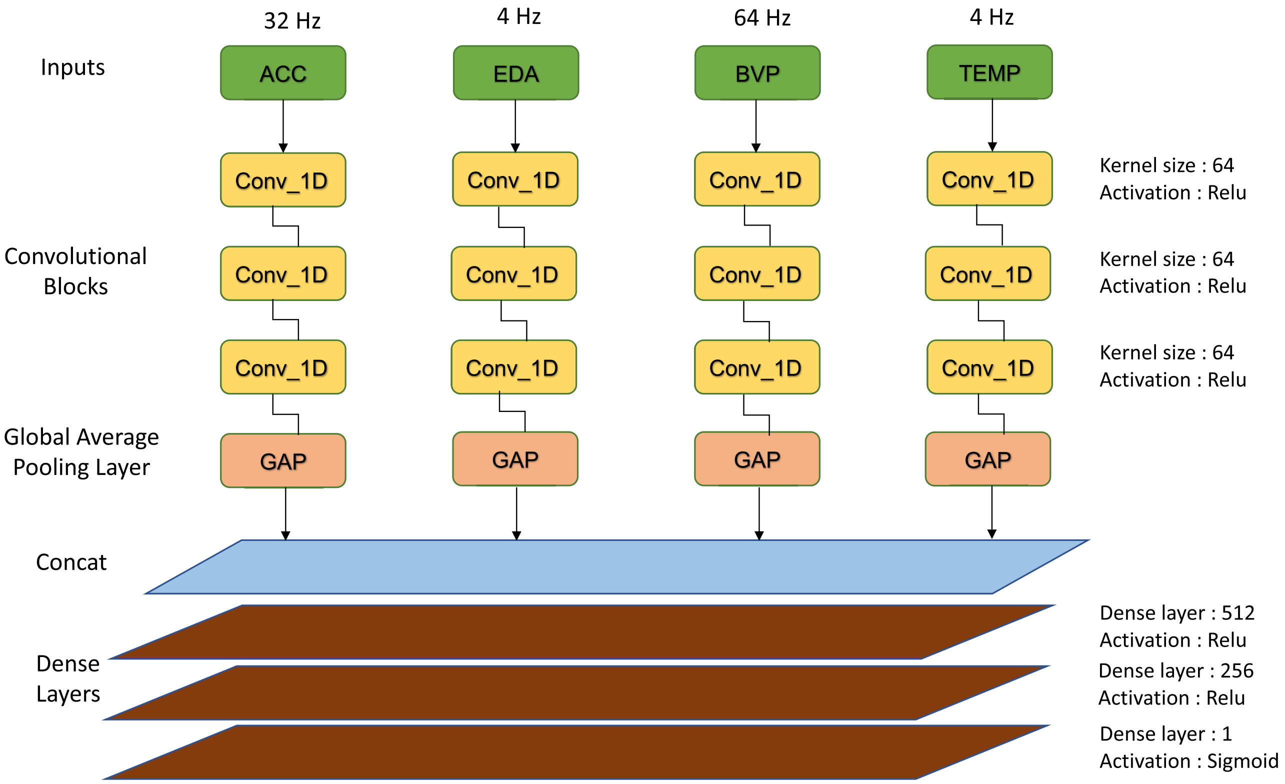


FIGURE 5. Multichannel fully convolutional network architecture (FCN) depicting the different layers and the hyperparameters used.

and identifies the ones that yield the highest score during validation. A list of parameters were passed as arguments to the classifiers. Additionally, it was used in different iterations of the cross validation to determine which parameters were most optimally used. A four-class and a binary classification task were performed based on the responses from the participants. In the four-class classification task, the labels ranged from ‘1’ to ‘4’, with ‘1’ representing the least interesting and ‘4’ representing the most interesting. For the binary classification, the labels ‘1’ and ‘2’ were combined and indicated as ‘not interested’, while the labels ‘3’ and ‘4’ were combined and indicated as ‘interested’.

B. DEEP LEARNING-BASED INTEREST DETECTION

The deep neural networks reduce the need for feature engineering as they can learn high-level features in the hidden layers, decreasing the complexity of the flow and the amount of labor required while simultaneously improving the likelihood of obtaining the relevant information, which is sometimes unavailable even to domain specialists. By integrating layers, the layer-based structure of the neural network allows the creation of a wide range of deep learning architectures. For time series classification based on multimodal physiological signals, a one-dimensional convolutional neural network turned out to be efficient in extracting relevant information from raw sensor signals, facilitating better classification [44].

In addition to the machine learning classifiers, a deep learning model based on 1D CNN [45], CNN-LSTM, and FCN [24], [46] has been implemented for interest detection. Contrary to the machine learning classifiers, the main advantage of using a deep learning model is that it can learn high-level features from raw sensor signals, and manual feature extraction is not required, which requires domain expertise. The deep learning models were trained from the raw sensor input of the eye tracker and the physiological signals from the E4 wristband to predict interest.

1) FCN

FCN is a variant of convolutional neural networks (CNNs) and is particularly suited for time series classification tasks because it can learn features that are invariant to translation and scaling. The network can learn patterns in the data that occur at different points and can recognize these patterns regardless of their location or scale.

The FCN model consists of three convolutional blocks for each signal followed by a Global Average Pooling layer and the branches are concatenated and fed to one or more fully connected dense layer. Each convolutional layer applies a set of filters to the input time series, and the output of each layer is passed through a non-linear activation function such as the

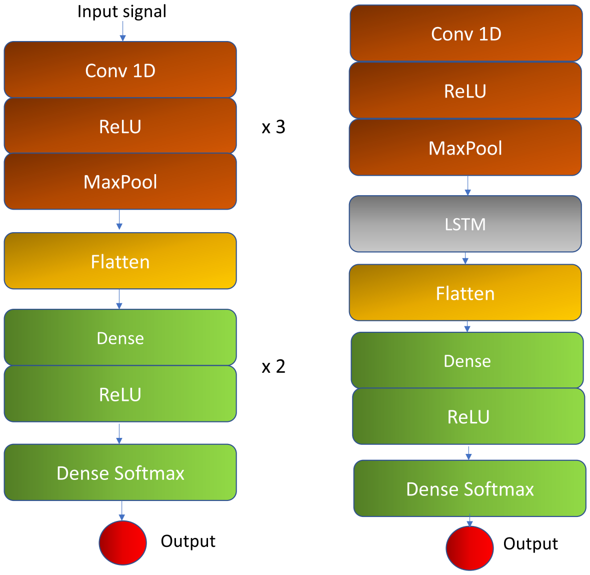


FIGURE 6. **1D CNN and CNN-LSTM architecture.**

rectified linear unit (ReLU) function. The fully connected layers then process the output of the convolutional layers to make the final classification. The advantage of using FCN for time series classification is that it can handle variable-length time series without the need for padding or truncation. The model architecture is shown in Figure 5.

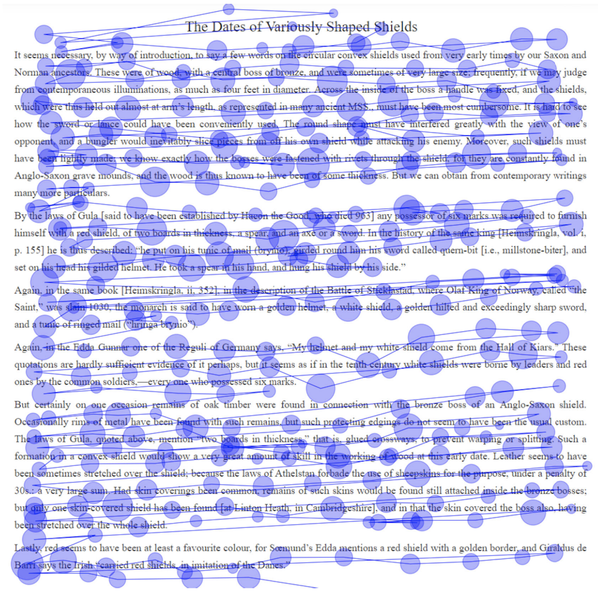
2) CNN-LSTM

CNN-LSTM is a type of neural network architecture that combines convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. The basic idea behind CNN-LSTM is to use CNNs to extract local features from the time series data and use LSTMs to capture long-term dependencies in the data.

The architecture of CNN-LSTM typically consists of a series of convolutional layers followed by one or more LSTM layers. The convolutional layers are used to extract local features from the time series data, which are then passed to the LSTM layers. The LSTM layers are used to capture long-term dependencies in the data and make the final classification. The advantage of using CNN-LSTM for time series classification is that it can handle both local and global patterns in the data. The convolutional layers are particularly effective at capturing local patterns such as spikes and dips, while the LSTM layers can capture global patterns such as trends and seasonality. The architecture of the CNN-LSTM model and the 1D CNN model is shown in the Figure 6

IV. EVALUATION

An experiment involving 13 university students reading 18 newspaper articles was conducted to compare the proposed approaches, investigate effective features, and discuss future challenges.



The Dates of Variously Shaped Shields

It seems necessary, by way of introduction, to say a few words on the circular convex shields used from very early times by our Saxon and Norman ancestors. These were of wood, with a central boss of bronze, and were sometimes of very large size; frequently, if we may judge from contemporaneous illuminations, as much as four feet in diameter. Across the inside of the boss a handle was fixed, and the shields, which were thus held out almost at arm's length, as represented in many ancient MSS., must have been most cumbersome. It is hard to see how the sword or lance could have been conveniently used. The round shape must have interfered greatly with the view of one's opponent, and a bungler would inevitably slice pieces from off his own shield while attacking his enemy. Moreover, such shields must have been lightly made; we know exactly how the bosses were fastened with rivets through the shield, for they are constantly found in Anglo-Saxon grave mounds, and the wood is thus known to have been of some thickness. But we can obtain from contemporary writings many more particulars.

By the laws of Gula [said to have been established by Hacon the Good, who died 963] any possessor of six marks was required to furnish himself with a red shield, of two boards in thickness, a spear, and an axe or a sword. In the history of the same king [Heimskringla, vol. i, p. 155] he is thus described: "he put on his tunic of mail (brynio), girded round him his sword called quern-bit [i.e., millstone-biter], and set on his head his gilded helmet. He took a spear in his hand, and hung his shield by his side."

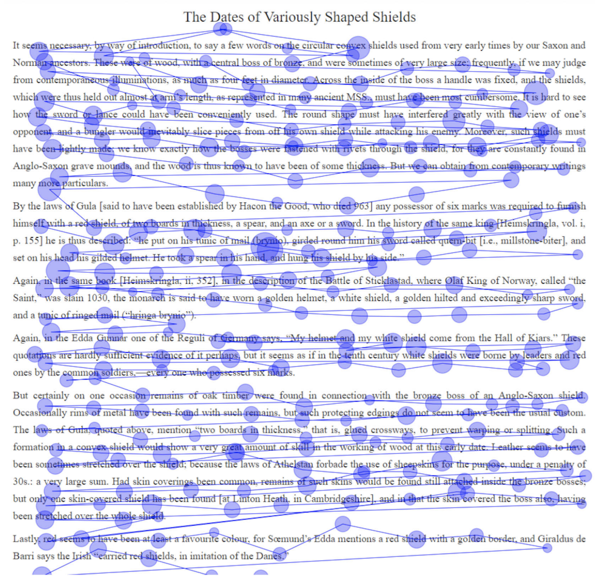
Again, in the same book [Heimskringla, ii, 352], in the description of the Battle of Sticklastad, where Olaf King of Norway, called "the Saint," was slain 1030, the monarch is said to have worn a golden helmet, a white shield, a golden hilted and exceedingly sharp sword, and a tunic of ringed mail ("hringa brynio").

Again, in the Edda Gunnar one of the Reguli of Germany says, "My helmet and my white shield come from the Hall of Kiars." These quotations are hardly sufficient evidence of it perhaps, but it seems as if in the tenth century white shields were borne by leaders and red ones by the common soldiers,—every one who possessed six marks.

But certainly on one occasion remains of oak timber were found in connection with the bronze boss of an Anglo-Saxon shield. Occasionally rims of metal have been found with such remains, but such protecting edgings do not seem to have been the usual custom. The laws of Gula, quoted above, mention "two boards in thickness," that is, glued crossways, to prevent warping or splitting. Such a formation in a convex shield would show a very great amount of skill in the working of wood at this early date. Leather seems to have been sometimes stretched over the shield; because the laws of Athelstan forbade the use of sheepskins for the purpose, under a penalty of 30s.: a very large sum. Had skin coverings been common, remains of such skins would be found still attached inside the bronze bosses; but only one skin-covered shield has been found [at Linton Heath, in Cambridgeshire], and in that the skin covered the boss also, having been stretched over the whole shield.

Lastly, red seems to have been at least a favourite colour, for Sœmund's Edda mentions a red shield with a golden border, and Giraldus de Barri says the Irish "carried red shields, in imitation of the Danes."

FIGURE 7. **A document that was rated as very interesting exhibiting gaze patterns characterized by a higher number and longer duration of fixations, along with increased reading time.**



The Dates of Variously Shaped Shields

It seems necessary, by way of introduction, to say a few words on the circular convex shields used from very early times by our Saxon and Norman ancestors. These were of wood, with a central boss of bronze, and were sometimes of very large size; frequently, if we may judge from contemporaneous illuminations, as much as four feet in diameter. Across the inside of the boss a handle was fixed, and the shields, which were thus held out almost at arm's length, as represented in many ancient MSS., must have been most cumbersome. It is hard to see how the sword or lance could have been conveniently used. The round shape must have interfered greatly with the view of one's opponent, and a bungler would inevitably slice pieces from off his own shield while attacking his enemy. Moreover, such shields must have been lightly made; we know exactly how the bosses were fastened with rivets through the shield, for they are constantly found in Anglo-Saxon grave mounds, and the wood is thus known to have been of some thickness. But we can obtain from contemporary writings many more particulars.

By the laws of Gula [said to have been established by Hacon the Good, who died 963] any possessor of six marks was required to furnish himself with a red shield, of two boards in thickness, a spear, and an axe or a sword. In the history of the same king [Heimskringla, vol. i, p. 155] he is thus described: "he put on his tunic of mail (brynio), girded round him his sword called quern-bit [i.e., millstone-biter], and set on his head his gilded helmet. He took a spear in his hand, and hung his shield by his side."

Again, in the same book [Heimskringla, ii, 352], in the description of the Battle of Sticklastad, where Olaf King of Norway, called "the Saint," was slain 1030, the monarch is said to have worn a golden helmet, a white shield, a golden hilted and exceedingly sharp sword, and a tunic of ringed mail ("hringa brynio").

Again, in the Edda Gunnar one of the Reguli of Germany says, "My helmet and my white shield come from the Hall of Kiars." These quotations are hardly sufficient evidence of it perhaps, but it seems as if in the tenth century white shields were borne by leaders and red ones by the common soldiers,—every one who possessed six marks.

But certainly on one occasion remains of oak timber were found in connection with the bronze boss of an Anglo-Saxon shield. Occasionally rims of metal have been found with such remains, but such protecting edgings do not seem to have been the usual custom. The laws of Gula, quoted above, mention "two boards in thickness," that is, glued crossways, to prevent warping or splitting. Such a formation in a convex shield would show a very great amount of skill in the working of wood at this early date. Leather seems to have been sometimes stretched over the shield; because the laws of Athelstan forbade the use of sheepskins for the purpose, under a penalty of 30s.: a very large sum. Had skin coverings been common, remains of such skins would be found still attached inside the bronze bosses; but only one skin-covered shield has been found [at Linton Heath, in Cambridgeshire], and in that the skin covered the boss also, having been stretched over the whole shield.

Lastly, red seems to have been at least a favourite colour, for Sœmund's Edda mentions a red shield with a golden border, and Giraldus de Barri says the Irish "carried red shields, in imitation of the Danes."

FIGURE 8. **A document that was rated as boring exhibiting gaze patterns characterized by a decreased number and shorter duration of fixations, accompanied by a reduction in reading time.**

A. EXPERIMENTAL DESIGN

Thirteen university students (mean age: 25, std: 3, male: 6, female: 7; two of them are familiar with eye-tracking) participated in the experiment where each of them were asked to read 20 newspaper articles comprising 403 × 649 words each (mean: 555, std: 70). The Figure 7 shows the gaze pattern of a document rated as very interesting by participant *P9*, and the figure 8 shows the gaze pattern of the same document rated as very boring by participant *P11*. The data from two articles were removed from all participants due to an error in data collection. The newspaper articles were selected from BBC news, as they seemed to better capture the purpose of the experiment than any other text. To capture the reader's interest, a wide range of topics from different platforms, such as technology, politics, sports, and cooking were obtained. Before the experiment, all participants were informed about the purpose of the study, and the experiment was carried out with their consent.

Table 3 presents the questionnaires administered to the participants after reading each article. The questions in the survey comprised of different aspects, including the participants' subjective comprehension (Q1), their level of interest in the article (serving as ground-truth) (Q2), and a question specifically related to the article aimed at assessing objective comprehension (Q3). The participants were required to rate both Q1 and Q2 on a scale ranging from 1 to 4, with 1 representing the lowest rating and 4 denoting the highest rating.

The recordings were made in two sessions of one hour each to avoid eye fatigue, and these were conducted in a controlled environment. The eye-tracker and the desktop were maintained in a stable position. The lighting of the room was set so as not to affect the gaze data (pupil diameter). The calibration was done after reading every document to avoid errors or shifts in the gaze points.

B. EVALUATION PROTOCOL

Three distinct cross-validation techniques were used to separate the data for training and testing. In the leave-one-participant-out cross-validation (LOPOCV) technique, the data from a single participant was set aside as the test set, while the data from all the remaining participants was used for training the model. This process was repeated iteratively for each participant, and the resulting accuracies were averaged to obtain the overall performance evaluation of the model. By evaluating the model's performance on multiple test sets that consist of data from different participants, a more robust assessment of the model's ability to generalize to unseen data could be obtained. The averaged accuracies across these iterations provides an overall performance evaluation that accounts for individual differences among participants.

The leave-one-document-out cross-validation (LODOCV) technique involves excluding the data of a single document from the training set and utilizing it as the test set. This process was repeated iteratively for each document in the study, similar to the LOPOCV approach. By employing LODOCV, it is possible to gain a more profound understanding of how well the model generalizes to unseen documents and obtain insights specific to the individual documents used in the study. The person-specific technique [47] was another strategy used for data splitting, in which data from a single participant (each document used as a test from the same participant) was split for training and testing. The same procedure was followed in other participants, and the average accuracy was derived from all participants. Given that each person's involvement or interest is different, this strategy offers a more personalized way for interest detection.

C. RESULTS

For the manual feature extraction-based approach, a random forest and an SVM model were trained to predict the interest of the participants, which is used as a baseline for comparison. The extracted features from the gaze data and the Empatica E4 were used to train the models, and the accuracy, f1-score, precision and recall were computed individually and by the combination of features. In terms of accuracy and f1-score, the SVM model outperformed the random forest model. The results of the feature extraction-based approach only included the better performing SVM model, while the Random Forest results were omitted, as they have been mentioned in our previous works [21], [22]. The target labels ranged from 1-4 where 1 represents 'not at all interested (0%)', 2 : 'some of it (30%)', 3 : 'most of it (60%)' and 4 : 'all of it (100%)'. The SVM achieved a classification accuracy of 41.5% and a f1-score of 39.4 when using LOPOCV. When using LODOCV, the SVM achieved an accuracy of 47.2% and a f1-score of 44.6.

In addition to the four-class classification, a binary classification was also performed by combining the ground truth labels *one* and *two* with being indicated as *not-interested* and *three* and *four* combined with being indicated as *interested*. For the binary classification, similar to the four-class classification, the SVM model with the combination of E4 features and eye movements achieved better classification accuracies of 68% and 71% in LOPOCV and LODOCV, respectively.

For the deep learning-based approach, a CNN and CNN-LSTM model using gaze features, a Fully Convolutional Neural Network (FCN) using physiological features, and the combination of CNN-LSTM and FCN network using the gaze and E4 data were implemented to detect the interest among the users. From Table 4 it was observed that the CNN-LSTM model using the gaze features outperformed the SVM in terms of accuracy for both four-class and binary classification with LOPOCV and LODOCV. The CNN-LSTM model, when applied with the LOPOCV technique, demonstrated a notable performance improvement of 14% for binary classification in terms of accuracy. Similarly, when employing the LODOCV method, there was a 10% improvement in accuracy. These improvements indicate the efficacy of the CNN-LSTM model and the effectiveness of the cross-validation approaches in enhancing the model's performance.

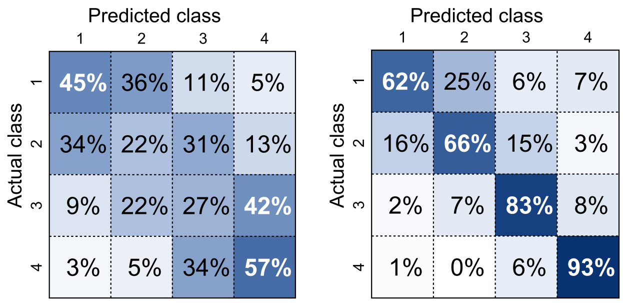
However, the combination of gaze and E4 features with CNN-LSTM and FCN models did not yield the expected results across all evaluation metrics. Despite the anticipation that incorporating gaze and E4 features would enhance the models' performance, they did not exhibit the desired impact on metrics such as accuracy, F1-score, or other evaluation measures. This suggests that the selected features were not adequately informative of effectively leveraging these particular feature combinations. A CNN model was also utilized

TABLE 3. The survey used for the experimental design.

No.	Questionnaire	Type	Range
Q1.	How much of the article did you understand?	Subjective comprehension	1-4 (1 being least and 4 highest)
Q2.	How much of the article did you find interesting?	Interest level	1-4 (1 being least and 4 highest)
Q3.	Question related to the article	Objective comprehension	Multiple choice

TABLE 4. Summary of classification results using machine learning and deep learning models for Leave One Participant Out (LOPO) and Leave One Document Out (LODO) cross-validation.

Models	Features	Validation	Binary				Four-class			
			Accuracy	F1-Score	Precision	Recall	Accuracy	F1-Score	Precision	Recall
SVM	Gaze + E4	LOPO	68	65	65.1	64.1	41.5	39.4	38.2	39.7
		LODO	71	70	68.7	69.2	47.2	44.6	44.2	43.7
FCN	E4	LOPO	60.8	59.2	59.1	59.8	28.3	27.8	27.1	28.3
		LODO	61.4	60.1	59.9	60.6	33.9	32.0	31.7	32.4
CNN-LSTM	**Gaze**	**LOPO**	**82.3**	**81.7**	**82.1**	**80.2**	**46.2**	**46.7**	**47.6**	**45.6**
		LODO	**80.5**	**80.2**	**80.4**	**79.6**	**52.8**	**51.8**	**52.4**	**50.5**
CNN-LSTM + FCN	Gaze + E4	LOPO	71.5	70.2	69.5	70.6	42.7	40.2	39.1	41.4
		LODO	70.9	69.3	69.1	69.2	46.1	45.9	45.2	46.4



(a) Participant-independent, acc.: 0.46 (b) Participant-dependent, acc.: 0.78

FIGURE 9. Confusion matrices for the four class interest level classification where labels are 1 : 'not at all interested (0%)', 2 : 'some of it (30%)', 3 : 'most of it (60%)' and 4 : 'all of it (100%).

in the experiment, although its prediction scores were lower compared to those of the CNN-LSTM model. As a result, it has not been included in the table that presented the evaluation metrics.

In addition to the two cross-validation techniques, which are person and document-independent, a participant-dependent cross-validation technique was also implemented since the interest levels could be unique and vary depending on the participants reading the text. In the person-dependent approach, each document from the same participant was left out for testing, and the average accuracy was estimated. Figure 9 shows confusion matrices as the comparison. As expected, it was observed that there was a significant increase in accuracy using this approach, as the results were uniquely dependent on the particular person and more personalized.

The Pearson correlation of the features for each participant with the level of interest was plotted, and it was observed that the level of subjective comprehension of a person has a high impact on the level of reading interest in Figure 12. Similarly, objective comprehension also had a positive correlation with the level of interest for each participant. The correlation between features and labels was observed to be comparatively trivial, although a higher correlation was expected. With a few exceptions, each participant's correlation varied considerably for each feature. This convinced us that physiological predictions are not just document-dependent, but also user-dependent. Pearson correlation was chosen as it aligned with the specific goals and context of our study. Furthermore, in our research area, Pearson correlation has commonly been employed in previous studies, enabling meaningful comparisons and ensuring consistency in the literature.

V. DISCUSSION

This section provides a comprehensive discussion of the contributions made by this study, furnishing a detailed analysis of their significance and impact. Additionally, it sheds light on the challenges encountered throughout the study and addresses the limitations that may have influenced the research outcomes.

A. DEEP LEARNING-BASED OR MANUAL FEATURE EXTRACTION-BASED APPROACH

In the previous two studies [21], [22], the interest prediction of readers was based on a manual feature extraction-based approach, such as SVM and Random forest. While the manual feature extraction approach has been widely used and provided initial insights into interest prediction, there was room for improvement in terms of prediction accuracy. The reliance on preselected features can limit the ability to capture all the complex and nuanced patterns that influence reader interest. Consequently, the accuracy of the predictions may not fully capture the subtleties and dynamics of readers' preferences. The reason for selecting SVM and Random forest for our analysis was that both are less prone to overfitting compared to other machine learning algorithms, making it useful when the available data is limited. Random forest also provides a measure of feature importance, which can be useful in feature selection or feature engineering. In this work, an interest detection method employing raw features with deep learning-based approaches was implemented, which yielded an increase in the efficiency of interest prediction. A comparison was performed between the classification results achieved using an SVM model and deep-learning models.

From the achieved results, it could be concluded that the predictions were significantly improved using just the raw gaze features through the deep learning-based approach (see Table 4). The CNN and CNN-LSTM based models were able to outperform the accuracy obtained using the SVM model, which used a combination of features from the gaze and E4. The primary aim of this study was to evaluate whether deep learning-based techniques could surpass manual feature extraction-based approaches in terms of prediction accuracy, and the findings indicated that deep learning-based models could be exceedingly beneficial in detecting cognitive states such as interest or engagement in a person for a reading task. The main benefit of utilizing deep learning models for interest prediction is to nullify the workload of extracting the most relevant features for classification purposes, which also requires high domain expertise. The labor involved in determining the most relevant features from the physiological and eye gaze data could be negated with the deep learning-based architecture.

B. CONTRIBUTIONS OF EACH SENSING MODALITY

The combination of sensing modalities from the eye-tracker and E4 wristband generated better prediction accuracy for the manual feature extraction-based approach than a single sensor, but the raw gaze data alone provided a considerable performance boost for the deep learning-based approach. The CNN-LSTM model with the raw gaze features significantly improved the classification accuracy, but intriguingly, the combination of the sensing modalities did not enhance the performance as compared to using only gaze features. The FCN model was trained using physiological signals from E4 and then concatenated with the CNN-LSTM model to compute the results by fusing both sensors. The prediction results from the sensor fusion utilizing deep learning models were only modestly improved compared to the accuracy of the SVM classification. The raw gaze features using deep learning could be more accurate predictors for cognitive state estimation in an individual. Even though the performance of the deep learning models with the combination of sensor signals was comparable to the manual feature extraction-based approach, the raw gaze features from the eye tracker using deep learning models proved to be a better choice for predicting interest.

Lin et al. investigated the impact of the classification results at the sensor and signal levels and removing the EDA signal reduced the accuracy and weighted F1 score for affect recognition, providing an explanation at the signal level [25]. Similarly to their work, the variability in model performance was examined by excluding each type of signal from the E4 wristband. Each modality was left out to detect interest, with the aim of understanding the importance of each physiological signal in predicting the interest among users. Similar to the outcomes obtained by Lin et al., it could be observed from Table 5 that the accuracy was lowest when the EDA signal was removed, but the omission of ACC and BVP had little to no impact on the model's performance.

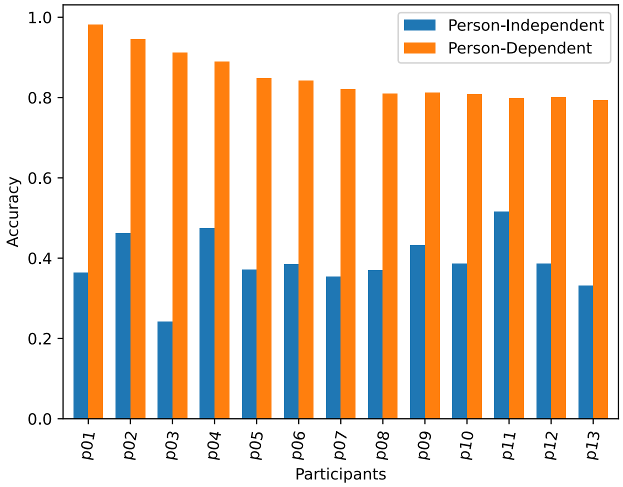


FIGURE 10. The plot showing the difference in accuracy per participant using a CNN-LSTM model with a person dependent and person independent approach.

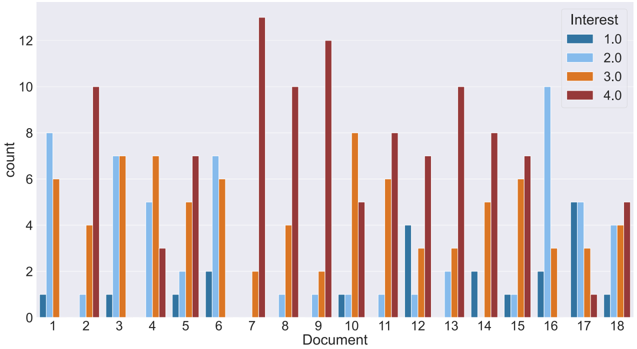
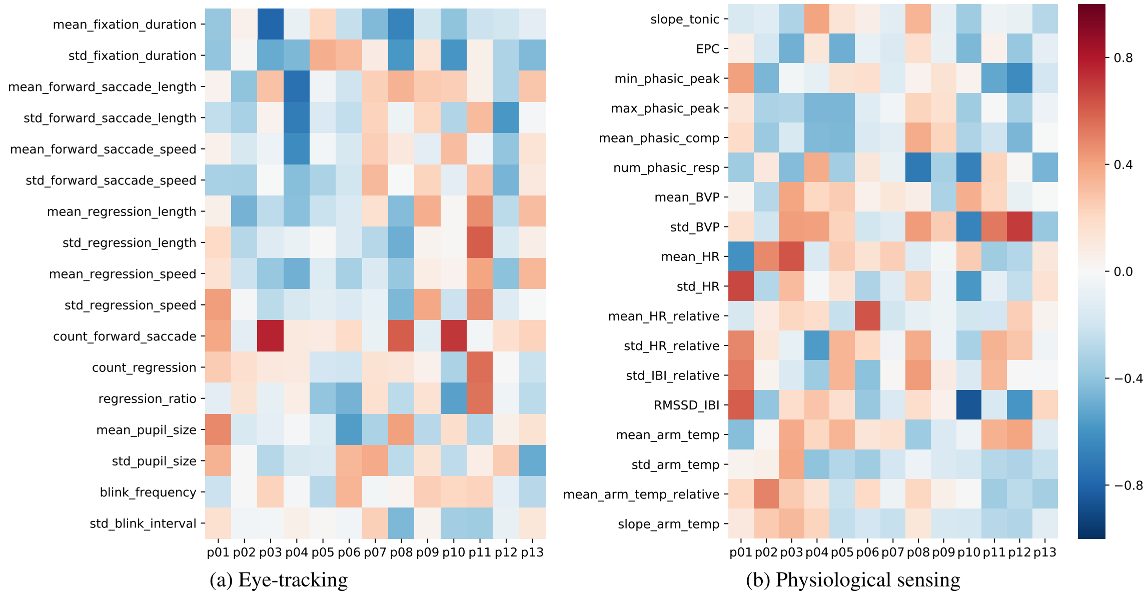


FIGURE 11. The interest ratings provided by the participants for all the documents used in the experiment.



(a) Eye-tracking (b) Physiological sensing

FIGURE 12. Pearson's correlations between interest and extracted features for each participant.

TABLE 5. The binary classification accuracy omitting each modality from Empatica E4. A decreased accuracy indicates the importance of the modality.

Modalities	Accuracy	F1-Score	Precision	Recall
Without ACC	58.77	56.50	57.24	55.65
Without BVP	58.17	54.89	54.32	55.14
Without Temp	56.78	55.26	56.73	54.29
Without EDA	**53.29**	**51.17**	**50.66**	**51.86**

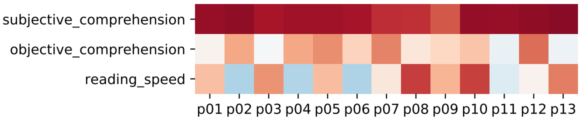


FIGURE 13. Pearson's correlations between interest and other surveys.

C. PERSON AND DOCUMENT DEPENDENCY

Depending on the context, a person's cognitive state can differ significantly, and each person has a distinct method for approaching a problem. To investigate how person-specific predictions compare against generalized models, interest was estimated at a person-specific level by using the data from a single participant as training and test. Every person's level of interest when reading a particular context can vary greatly, and what captures one individual's attention might not have the same effect on someone else.

To address this individuality, a CNN-LSTM model with raw gaze features was developed and implemented using a person-specific approach. This person-specific approach involved adapting the model to each individual's unique characteristics and preferences. By considering these individual factors, the model significantly outperformed the generalized models in accurately detecting interest levels. In fact, the person-specific CNN-LSTM model demonstrated a remarkable performance improvement of nearly 50% per participant compared to the other models. To illustrate the impact of this person-specific approach on interest detection, Figure 10 showcases the variance in accuracy among different participants. The plot provides a visual representation of the accuracy scores for each individual, highlighting how the person-specific approach influenced the accuracy levels in detecting interest. It is possible to observe variations in accuracy across the participants, which further emphasize the importance of considering individual characteristics and preferences in interest detection.

The reading experiment included a diverse range of materials, covering various categories such as technology, politics, sports, business, history, food, and more. Each participant engaged with these materials, and their responses varied significantly based on their individual level of interest. The experiment aimed to estimate the participants' subjective perceptions of interest in relation to the different documents provided.

The selection of newspaper articles from BBC News was motivated by their perceived ability to align closely with the purpose of the experiment. By choosing articles from this source, the aim was to ensure a certain level of credibility, accuracy, and depth in the information provided. To further engage the readers and maintain their interest throughout the experiment, the importance of incorporating a diverse range of topics was recognized. In order to achieve this, the search was expanded beyond a single platform, and articles were gathered from different sources and platforms. This approach allowed us to capture a broader spectrum of subjects that appeal to a wide range of readers.

By including articles from various domains such as technology, politics, sports, cooking, and more, the aim was to provide a rich and varied dataset that would cater to the diverse interests of readers. Additionally, an attempt was made to mitigate any potential biases associated with a single news source by incorporating articles from multiple platforms. It should be noted that different platforms often have their own editorial preferences, writing styles, and perspectives, which can introduce biases into the dataset. The aim was to create a more balanced and representative collection of texts by incorporating articles from various sources.

Figure 11 serves as a reference in assessing the participants' ratings for each document. Notably, document number 7 received consistently high ratings, indicating that it was perceived as highly interesting by a majority of the participants. Conversely, document number 16 received the lowest ratings and was considered the least interesting among the participants. An additional analysis of the data revealed interesting insights regarding the categories of the documents. It was observed that the document with the highest number of votes for being highly interesting belonged to the *technology* category. This suggests that the participants found the technological content engaging and captivating. On the other hand, the document that received the lowest number of votes, indicating the least interesting response, belonged to the *history* category. This suggests that the historical content failed to generate significant interest among the participants.

Due to the varying responses elicited by diverse documents, a comprehensive cross-validation procedure was implemented, employing the Leave-One-Document-Out Cross-Validation (LODOCV) approach. This methodology enabled a rigorous assessment and verification of the models employed in the experiment, thereby guaranteeing the reliability and robustness of the outcomes.

The correlation plot depicted in Figure 12 provides insights into the relationship between hand-crafted features and the level of interest. It indicates that there were significant variations in the correlation between gaze features and interest level across individual participants. This observation led to the conclusion that physiological predictions, derived from gaze features, are more reliant on the characteristics of individual users rather than the content of the documents themselves. In other words, the correlation between gaze features and interest is more user-dependent than document-dependent.

Figure 13 presents the Pearson correlations between the level of interest and other measures that were not included in the proposed methods. Notably, there is a high correlation between comprehension of a document and interest in reading. This finding aligns with our intuition that interest can only be experienced if the person understands the text. It reinforces the notion that comprehension plays a vital role in the perception and engagement of interest during reading. While the integration of user actions into the estimation task may have limited application scenarios, Figure 13 demonstrates that incorporating user actions improves the performance of interest estimation. This indicates that additional user actions, beyond gaze features alone, can contribute to a more accurate assessment of interest levels.

D. LIMITATIONS

Even though the deep learning-based approach provides more flexibility and is more effective in cognitive state prediction, there are certain limitations like the lack of data, expensive training due to complex data models, and high demand for computational power. For the evaluation, the data was collected from 13 participants reading 18 different documents, and that might not be enough to train complex deep learning models. It also requires extensive hardware for performing complex mathematical calculations. The Empatica E4 recorded data at lower sampling frequencies, which could explain why the FCN model did not function well with the physiological data as expected. Initially, the expectation was that the raw physiological features recorded from the E4 wristband would yield better prediction results when combined with a deep learning-based approach. However, the actual results were comparable to those achieved using a manual feature extraction-based approach. This suggests that the deep learning model's performance did not significantly benefit from the raw physiological features, indicating that the relationship between these features and interest was not effectively captured by the model. For future work, the plan is to use more sophisticated sensors with higher sampling frequencies that could facilitate cognitive state detection using deep learning models.

The integration of gaze and E4 features with CNN-LSTM and FCN models did not produce the anticipated outcomes across all evaluation metrics. Further investigation and analysis are required to understand the reasons behind the suboptimal performance of the combined gaze and E4 features with CNN-LSTM and FCN models. The combination of gaze and E4 features was anticipated to provide complementary information, with gaze features capturing visual attention and E4 features capturing physiological signals. However, the unexpected results suggest that the information provided by these features might not be as complementary as initially assumed. This could indicate that the features are capturing redundant or irrelevant information, hindering the models from effectively leveraging their combined power. The quality and reliability of the gaze and E4 features are also crucial factors to consider. The combined feature approach could have been influenced by issues such as noise, outliers, or inconsistencies in the collected data. Likewise, it is possible that the chosen model architectures or hyperparameter settings were not optimal for effectively integrating the gaze and E4 features. Given the unexpected results, future directions involve investigating alternative feature combinations, exploring different model architectures, or considering additional relevant factors that may enhance the performance of interest detection models.

The selection of thirteen university students as participants in the experiment may indeed be considered a limited and potentially biased sample. While the study provides some valuable insights, the small sample size of thirteen participants limits the generalizability of the findings to a broader population. As university students represent a specific demographic group with unique characteristics, the results may not be representative of the general population. Different age groups, educational backgrounds, and cultural factors could influence reading behaviors and eye-tracking patterns. Despite our attempt to maintain a balanced gender distribution in our experiment (with six male and seven female participants), the unequal representation of male and female participants poses challenges in drawing conclusions that can be uniformly applied to both genders.

The extensive evidence supports that gender influences reading strategies, comprehension, and eye movement patterns, thus emphasizing the importance of equal gender representation in research studies. Additionally, the fact that only two out of the thirteen participants were familiar with eye-tracking technology introduces another potential bias. Participants with prior experience using eye-tracking devices may exhibit different eye movement patterns and reading behaviors compared to those who are unfamiliar with the technology. This familiarity could impact the accuracy and reliability of the eye-tracking data, and subsequently influence the classifications derived from it. To address these limitations, future studies could benefit from a more diverse participant pool that includes individuals from various age groups, educational backgrounds, and gender identities, which would help to improve the external validity and enhance the generalizability of the research findings.

Another limitation of our work was the participant leniency in the responses collected as ground truth. The articles collected for the experiment were carefully segmented to induce varying levels of interest in the participants, but more participants tended to rate the documents as *highly interesting*, which led to an imbalance in the class distribution. For future work, the plan is to implement a system with preset conditions that could reduce the disparity in class distributions and ensure better ground truth collection.

VI. CONCLUSION AND FUTURE WORK

The study aimed to measure the level of interest experienced by users while reading newspaper articles by integrating gaze data and physiological responses. A multimodal approach was adopted, combining gaze data and physiological measurements, to provide a comprehensive and accurate assessment of interest levels. The experiment involved 13 university students reading 18 newspaper articles obtained from the BBC news database. To predict the participants' interest levels, two approaches were employed: a manual feature extraction-based approach and a deep learning-based approach. The outcomes demonstrated that the deep learning-based approach is more effective than the manual feature extraction-based approach in predicting the cognitive state of an individual, such as interest or engagement. A CNN-LSTM based model with the raw gaze features achieved the highest accuracy of 52% for four-class classification and 82.3% accuracy for binary classification. A person-dependent approach was employed for more personalized predictions and to estimate the interest level specific to an individual, and the accuracy per participant in this approach was observed to be better than the generalized approach. To assess the significance of physiological signals in detecting interest, each modality of the Empatica E4 wristband was omitted, and it was observed that the EDA signal is significantly relevant in interest detection. These findings contribute to the understanding of interest in the context of reading and offer potential applications for personalized learning experiences, enhanced comprehension, and motivation in educational settings. Further research in this area can explore ways to enhance the assessment of interest and its impact on learning outcomes.

The future tasks involve creating a system capable of identifying interest or motivation levels during reading in real-time, using gaze and physiological data, and implementing interventions in the form of feedbacks, alerts, or notifications that can improve reading skills and optimize the reading environment. In the future, it may be possible to personalize text for readers with greater accuracy, considering their individual reading preferences and presenting them with more engaging data. This could involve augmenting and anticipating the user's thought processes, as well as taking into account their state of mind and level of fatigue. Furthermore, this research could be expanded to create a teaching assistant that would provide teachers with more profound insights into their students' learning and understanding. Ultimately, this could lead to better designed articles and a more satisfying human-computer interaction experience.

REFERENCES

[1] A. Krapp, "The construct of interest: Characteristics of indvidual interests and interest–related actions from the perspective of a person–object–theory," in *Studies in Educational Psychology*. Munich, Germany: Inst. für Erziehungswiss. Und Pädag. Psychologie, Univ. der Bundeswehr, 1993.

[2] A. Krapp, "Interest, learning and academic achievement. Paper prepared for the symposium on' task motivation by interest," in *Proc. 3rd Eur. Conf. Learn. Instruct. (EARLI)*, Madrid, Spain, Sep. 1989, pp. S.299–S.323.

[3] J. I. Rotgans and H. G. Schmidt, "Situational interest and academic achievement in the active-learning classroom," *Learn. Instruct.*, vol. 21, no. 1, pp. 58–67, Feb. 2011.

[4] J. M. Harackiewicz and C. S. Hulleman, "The importance of interest: The role of achievement goals and task values in promoting the development of interest," *Social Personality Psychol. Compass*, vol. 4, no. 1, pp. 42–52, Feb. 2010.

[5] P. J. Silvia, "Interest—The curious emotion," *Current Directions Psychol. Sci.*, vol. 17, no. 1, pp. 57–60, Feb. 2008.

[6] U. Schiefele, "Interest, learning, and motivation," *Educ. Psychologist*, vol. 26, no. 3, pp. 299–323, Jun. 1991.

[7] S. Hidi, "Interest: A unique motivational variable," *Educ. Res. Rev.*, vol. 1, no. 2, pp. 69–82, Jan. 2006.

[8] K. A. Renninger and R. H. Wozniak, "Effect of interest on attentional shift, recognition, and recall in young children," *Develop. Psychol.*, vol. 21, no. 4, pp. 624–632, Jul. 1985.

[9] D. H. Hockenbury and S. E. Hockenbury, *Discovering Psychology*. Derbyshire, U.K.: Worth Publishers, 2006.

[10] C. E. Izard, *Human Emotions*. New York, NY, USA: Plenum Press, 1977.

[11] C. E. Izard, "Basic emotions, relations among emotions, and emotion-cognition relations," *Psychol. Rev.*, vol. 99, no. 3, pp. 561–565, 1992.

[12] W. L. Libby, B. C. Lacey, and J. I. Lacey, "Pupillary and cardiac activity during visual attention," *Psychophysiology*, vol. 10, no. 3, pp. 270–294, May 1973.

[13] S.-N. Yang and G. Mcconkie, ''Eye movements during reading: A theory of saccade initiation times,'' *Vis. Res.*, vol. 41, pp. 3567–3585, Feb. 2001.

[14] K. Rayner, K. H. Chace, T. J. Slattery, and J. Ashby, ''Eye movements as reflections of comprehension processes in reading,'' *Sci. Stud. Reading*, vol. 10, no. 3, pp. 241–255, Jul. 2006.

[15] R. Biedert, G. Buscher, S. Schwarz, M. Möller, A. Dengel, and T. Lottermann, ''The text 2.0 framework: Writing web-based gaze-controlled realtime applications quickly and easily,'' in *Proc. Workshop Eye Gaze Intell. Human Mach. Interact.*, Feb. 2010, pp. 114–117.

[16] A. Greco, A. Lanata, L. Citi, N. Vanello, G. Valenza, and E. P. Scilingo, ''Skin admittance measurement for emotion recognition: A study over frequency sweep,'' *Electronics*, vol. 5, no. 3, pp. 1–3, Aug. 2016.

[17] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, ''CvxEDA: A convex optimization approach to electrodermal activity processing,'' *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 797–804, Apr. 2016.

[18] S. Ishimaru, S. Bukhari, C. Heisel, N. Großmann, P. Klein, J. Kuhn, and A. Dengel, ''Augmented learning on anticipating textbooks with eye tracking,'' in *Positive Learning in the Age of Information*. Wiesbaden, Germany: Springer, 2018, pp. 387–398.

[19] D. Girardi, N. Novielli, D. Fucci, and F. Lanubile, ''Recognizing developers' emotions while programming,'' in *Proc. IEEE/ACM 42nd Int. Conf. Softw. Eng. (ICSE)*, Oct. 2020, pp. 666–677.

[20] I. Brishtel, A. A. Khan, T. Schmidt, T. Dingler, S. Ishimaru, and A. Dengel, ''Mind wandering in a multimodal reading setting: Behavior analysis & automatic detection using eye-tracking and an EDA sensor,'' *Sensors*, vol. 20, no. 9, p. 2546, Apr. 2020.

[21] S. Jacob, S. Ishimaru, S. S. Bukhari, and A. Dengel, ''Gaze-based interest detection on newspaper articles,'' in *Proc. 7th Workshop Pervasive Eye Tracking Mobile Eye-Based Interact.*, Jun. 2018, pp. 1–7.

[22] S. Jacob, S. Ishimaru, and A. Dengel, ''Interest detection while reading newspaper articles by utilizing a physiological sensing wristband,'' in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, New York, NY, USA, Oct. 2018, pp. 78–81.

[23] R. Mutegeki and D. S. Han, ''A CNN-LSTM approach to human activity recognition,'' in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Feb. 2020, pp. 362–366.

[24] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[25] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, ''An explainable deep fusion network for affect recognition using physiological signals,'' in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 2069–2072.

[26] A. Abedin, F. Motlagh, Q. Shi, H. Rezatofighi, and D. Ranasinghe, ''Towards deep clustering of human activities from wearables,'' in *Proc. Int. Symp. Wearable Comput.*, Sep. 2020, pp. 1–6.

[27] P. Bobade and M. Vani, ''Stress detection with machine learning and deep learning using multimodal physiological data,'' in *Proc. 2nd Int. Conf. Inventive Res. Comput. Appl. (ICIRCA)*, Jul. 2020, pp. 51–57.

[28] F. Al Machot, A. Elmachot, M. Ali, E. A. Machot, and K. Kyamakya, ''A deep-learning model for subject-independent human emotion recognition using electrodermal activity sensors,'' *Sensors*, vol. 19, no. 7, p. 1659, Apr. 2019.

[29] F. Ordóñez and D. Roggen, ''Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition,'' *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016.

[30] J. Suzuki, H. Nittono, and T. Hori, ''Level of interest in video clips modulates event-related potentials to auditory probes,'' *Int. J. Psychophysiol.*, vol. 55, no. 1, pp. 35–43, Jan. 2005.

[31] Y. Shigemitsu and H. Nittono, ''Poster: Assessing interest level during movie watching with brain potentials,'' in *Proc. 2nd Int. Workshop Kansei*, Fukuoka, Japan, 2008, pp. 39–42.

[32] J. Azcarraga and M. T. Suarez, ''Recognizing student emotions using brainwaves and mouse behavior data,'' *Int. J. Distance Educ. Technol.*, vol. 11, no. 2, pp. 1–15, Apr. 2013.

[33] G. Vecchiato, A. G. Maglione, P. Cherubino, B. Wasikowska, A. Wawrzyniak, A. Latuszynska, M. Latuszynska, K. Nermend, I. Graziani, M. R. Leucci, A. Trettel, and F. Babiloni, ''Neurophysiological tools to investigate consumer's gender differences during the observation of TV commercials,'' *Comput. Math. Methods Med.*, vol. 2014, pp. 1–12, Jan. 2014.

[34] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias, ''Estimation of behavioral user state based on eye gaze and head pose—Application in an e-learning environment,'' *Multimedia Tools Appl.*, vol. 41, no. 3, pp. 469–493, Feb. 2009.

[35] S. Mota and R. W. Picard, ''Automated posture analysis for detecting Learner's interest level,'' in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2003, p. 49.

[36] A. Kapoor and R. W. Picard, ''Multimodal affect recognition in learning environments,'' in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, New York, NY, USA, Nov. 2005, pp. 677–682.

[37] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson, ''Emotion sensors go to school,'' in *Proc. Conf. Artif. Intell. Educ., Building Learn. Syst. That Care, Knowl. Represent. Affect. Modelling*, 2009, pp. 17–24.

[38] S. Ishimaru, S. Jacob, A. Roy, S. S. Bukhari, C. Heisel, N. Großmann, M. Thees, J. Kuhn, and A. Dengel, ''Cognitive state measurement on learning materials by utilizing eye tracker and thermal camera,'' in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 8, Nov. 2017, pp. 32–36.

[39] A. Babiker, Y. Baashar, A. A. Alkahtani, I. Faye, and G. Alkawsi, ''Towards detection of interest using physiological sensors,'' *Appl. Sci.*, vol. 11, no. 3, p. 1318, Feb. 2021.

[40] S. Salcedo-Sanz, J. L. Rojo-Álvarez, M. Martínez-Ramón, and G. Camps-Valls, ''Support vector machines in engineering: An overview,'' *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 4, no. 3, pp. 234–267, May 2014.

[41] R. Bednarik, S. Eivazi, and H. Vrzakova, ''A computational approach for prediction of problem-solving behavior using support vector machines and eye-tracking data,'' in *Eye Gaze in Intelligent User Interfaces*. London, U.K.: Springer, 2013, pp. 111–134.

[42] M. Pal, ''Random forest classifier for remote sensing classification,'' *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, Jan. 2005.

[43] M. Belgiu and L. Drăguţ, ''Random forest in remote sensing: A review of applications and future directions,'' *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 24–31, Apr. 2016.

[44] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Müller, ''Deep learning for time series classification: A review,'' *Data Mining Knowl. Discovery*, vol. 33, no. 4, pp. 917–963, Mar. 2019.

[45] D. Zhang, D. Cao, and H. Chen, ''Deep learning decoding of mental state in non-invasive brain computer interface,'' in *Proc. Int. Conf. Artif. Intell., Inf. Process. Cloud Comput.*, Dec. 2019, pp. 1–5.

[46] M. Dzieüyc, M. Gjoreski, P. Kazienko, S. Saganowski, and M. Gams, ''Can we ditch feature engineering? End-to-end deep learning for affect recognition from physiological sensor data,'' *Sensors*, vol. 20, no. 22, p. 6535, Nov. 2020.

[47] K. Nkurikiyeyezu, A. Yokokubo, and G. Lopez, ''The effect of person-specific biometrics in improving generic stress predictive models,'' 2019, *arXiv:1910.01770*.

JAYASANKAR SANTHOSH was born in Kerala, India, in 1992. He received the bachelor's degree in computer science from Mahatma Gandhi University, India, in 2014, and the master's degree in computer science from Technical University Kaiserslautern, Germany, in 2018.

From 2017 to 2019, he was a Research Assistant with the German Research Center for Artificial Intelligence (DFKI). Since 2019, he has been a Ph.D. Researcher with DFKI, where he is currently a member of the Immersive Quantified Learning Laboratory (IQL Lab) and the Smart Data and Knowledge Services (SDS) Department. Since 2022, he has also been a Teaching Assistant with the University of Kaiserslautern-Landau. He has published papers at Ubicomp and IUI conferences. His research interests include deep learning-based affective state recognition, assessing student involvement in e-learning, sensor data analysis, and feedback-based intervention in e-learning. He has been a professional member of the Association for Computing Machinery (ACM).

DAVID DZSOTJAN was born in Esztergom, Hungary, in 1982. He received the Diploma degree in physicist-engineering from the Budapest University of Technology and Economics, in 2006, and the Ph.D. degree in theoretical physics from the University of Kaiserslautern, Kaiserslautern, Germany, in 2011, with a focus on nanotechnology and quantum computing.

From 2012 and 2020, he was involved in research, mathematically modeling, and numerically simulating metallic nanostructures with the Wigner Research Center for Physics, Budapest, Hungary; Université de Bourgogne Franche-Comté, Dijon, France; and the University of Kaiserslautern. He is currently a Postdoctoral Researcher with the University of Kaiserslautern Landau and an External Researcher with the German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, and LMU, Munich, Germany. Since 2020, he has shifted the focus of studies to science education research, conceiving and investigating novel, AI-powered methods, and learning tools for understanding concepts in the STEM disciplines. His research interest includes assessing the learning gain and interest through eye-tracking and physiological signals in diverse learning environments.

SHOYA ISHIMARU (Member, IEEE) was born in Ehime, Japan, in 1991. He received the B.E. degree in electrical engineering and the M.E. degree in information science from Osaka Prefecture University, Japan, in 2014 and 2016, respectively, and the Ph.D. degree (summa cum laude) in engineering from the University of Kaiserslautern, Germany, in 2019.

From 2016 to 2019, he was a Research Associate with the University of Kaiserslautern and a Researcher with the German Research Center for Artificial Intelligence (DFKI), Germany. Since 2014, he has been a Researcher with the Keio Media Design Research Institute. Since 2021, he has also been a Junior Professor with the Department of Computer Science, University of Kaiserslautern-Landau (appointed with the University of Kaiserslautern, in 2021, and its campus was merged with the University of Koblenz and Landau, in 2023). Since 2022, he has been a Visiting Associate Professor with Osaka Metropolitan University. In addition, he has been a Senior Researcher with DFKI. His research interests include human–computer interaction, machine learning, and cognitive psychology toward amplifying human intelligence.

Prof. Ishimaru's awards and honors include the Best Presentation Award by the Asian CHI Symposium 2020, the Poster Track Honorable Mention by UbiComp/ISWC 2018, and the MITOU Super Creator which is a title given to outstanding software developers (around ten people per year) by the Ministry of Economy, Trade, and Industry, Japan.