IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 November 2023, accepted 28 December 2023, date of publication 8 January 2024,
date of current version 16 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3350745

RESEARCH ARTICLE

Toward an Interactive Reading Experience: Deep Learning Insights and Visual Narratives of Engagement and Emotion

JAYASANKAR SANTHOSH[1,2], (Member, IEEE), AKSHAY PALIMAR PAI[1], AND SHOYA ISHIMARU[2,3], (Member, IEEE)

[1]Department of Computer Science, University of Kaiserslautern-Landau, 67663 Kaiserslautern, Germany
[2]German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany
[3]Department of Computer Science, Osaka Metropolitan University, Osaka 558-8585, Japan

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

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 Engagement and emotion are critical components that significantly influence a reader's experience during a reading task. Despite the crucial role of engagement and emotions in shaping our reading experience, accurately tracking these dynamic states during actual reading remains a significant challenge. This study bridges this gap by detecting engagement and emotion levels during a reading task by leveraging the power of state-of-the-art deep learning models and investigating the correlations between the engagement levels and emotions. An experiment was conducted involving 18 university students reading 14 documents followed by a questionnaire to rate their levels of engagement, valence, and arousal after reading each document. A Tobii 4C eye-tracker with a pro license along with an Empatica E4 wristband were utilized to record behavioral and physiological data from the participants. A range of deep learning models were utilized for computing the engagement, valence, and arousal values, employing both user-independent and user-dependent methods. Our investigation revealed distinct yet complementary strengths in two deep learning models: Transformer excelled in user-independent detection of engagement and emotion with an accuracy of 80.38% (engagement), 71.28% (arousal) and 73.98% (valence) while ResNet shined in the user-dependent setting with an accuracy of 93.56% (engagement), 90.62% (arousal) and 88.70% (valence) which highlights the interplay between individual differences and reading dynamics. Intriguingly, we observed strong, document-specific correlations between engagement and emotion states, suggesting that different texts evoke unique affective responses. We developed an interactive dashboard visualizing predicted engagement and emotions, offering real-time feedback and personalized learning possibilities. The dashboard features an engagement gauge that displays the reader's level of engagement based on predicted class probabilities, and an emotion emoji serving as a visual cue that illustrates the predicted emotional state of the reader. This technology can inform the design of dynamic interfaces that adjust to individual reading styles and emotional responses, potentially enhancing comprehension and involvement.

INDEX TERMS Digital reading, physiological sensing, eye tracking, deep learning, affective state.

I. INTRODUCTION

Understanding the dynamics of engagement and emotion during reading tasks is crucial for improving comprehension and user experiences in various domains, including education, content creation, and human-computer interaction [1]. The degree of engagement reflects the reader's level of immersion and attentiveness to the content, while emotion captures the array of affective responses elicited by the text. A good understanding of these elements is important in many

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

FIGURE 1. The design overview.

areas, such as education, marketing, and user experience design [2], [3].

In the context of an e-learning environment, the ability to accurately determine students' engagement levels is of utmost importance. Engagement is a key indicator of students' active involvement and motivation in the learning process, directly impacting their academic performance and overall learning outcomes [4], [5]. To enhance the efficacy of e-learning platforms, it becomes crucial to leverage students' emotions as a means to evaluate their engagement effectively [6]. However, the current scenario presents several challenges that impede the integration of engagement and emotion analysis in the e-learning environment. One primary obstacle lies in the unavailability of large amounts of recorded data that capture students' engagement and emotional responses during the learning process. The scarcity of such data hinders the development of robust machine learning models capable of accurately predicting engagement levels based on emotional cues.

The prediction of a reader's internal states, including their level of engagement, holds the potential to play a significant role in enhancing the interactivity of reading material [7]. Ishimaru et al. [7] hypothesized that aligning presented information with students' current cognitive state would potentially enhance their learning and comprehension abilities. By gaining insights into the reader's emotional and cognitive responses, content creators and educators can tailor the reading experience to be more engaging and personalized. The emotional aspect of engagement is often overlooked in existing e-learning systems. While traditional metrics may provide insights into students' participation and completion rates, they fail to comprehend the underlying emotional experiences that drive these actions. Considering only quantitative metrics without examining the emotional dimension may lead to an incomplete understanding of students' true engagement levels and may hinder the potential for personalized learning experiences.

Current emotion detection techniques tend to be more oriented towards image data analysis rather than physiological data analysis [8], [9]. While image-based emotion detection has shown promise in various applications, it may not fully capture the intricacies of emotional responses during the learning process. Physiological data analysis, on the other hand, holds immense potential for uncovering subtle emotional fluctuations, such as changes in heart rate, skin conductance, or facial expressions, which are closely linked to students' engagement and cognitive processes [10], [11].

In addition to physiological signals, eye-tracking data also plays a pivotal role in detecting engagement and emotion during reading tasks [12], [13]. By tracking the direction and movement of a reader's eyes as they interact with the text, gaze data provides valuable insights into their attention, focus, and cognitive processing. Gaze data can reveal emotional responses, as increased fixations on emotionally charged words or sentences may indicate heightened emotional involvement. Combining gaze data with machine learning and analytical techniques enables researchers to develop models that accurately detect engagement and emotion, providing a deeper understanding of readers' cognitive and emotional states while reading [14].

The combination of gaze data and physiological signals offers distinct advantages in detecting emotion and engagement during reading tasks [15]. Brishtel et al. [15] demonstrated the high potential of the EDA sensor as a tool for detecting mind wandering, and integrating electrodermal activity with eye-tracking features notably enhanced the accuracy of mind wandering classification. Gaze data provides valuable insights into the reader's visual attention and focus on specific elements of the text, while physiological signals, such as heart rate and skin conductance, offer objective measures of emotional arousal and valence [16], [17]. By integrating these modalities, a more comprehensive understanding of the reader's cognitive and emotional states can be achieved, resulting in increased accuracy and reliability of emotion and engagement detection. This multimodal approach enables real-time assessment, facilitating personalized and adaptive reading experiences based on the reader's emotional responses. Moreover, the unobtrusive data collection process ensures a natural reading environment, enhancing the generalizability and applicability of findings across diverse populations and reading contexts.

Over the past decade, deep learning (DL) has emerged as a pivotal force within the realm of artificial intelligence (AI), facilitating significant advancements and groundbreaking accomplishments in diverse domains, particularly observed in three major areas like image processing [18], [19], [20], natural language processing (NLP) [21], [22], and reinforcement learning [23], [24], [25]. Time series sensor

data, characterized by sequences of measurements taken at successive points, presents unique challenges that make the application of deep learning more complex. Traditional deep learning models, highly effective in other domains, might not be directly transferable to the analysis of time series sensor data. The scarcity of deep learning methods in this area might be attributed to several factors, such as the complexity of temporal dependencies, the need for robust preprocessing techniques, the high dimensionality of data, and the varying sampling frequencies. Furthermore, the design and customization of deep learning architectures that are specifically tailored to capture the underlying patterns and dependencies in time series data have not been thoroughly investigated.

One motivation behind this work was to explore the role and connection of engagement and elicited emotions separately in a student learning environment. Earlier studies have demonstrated that content evoking positive emotions, such as joy or inspiration, frequently results in increased engagement and conversely, content eliciting negative emotions, like sadness or boredom, often leads to decreased engagement. Rosa and Bernardo [26] found that a notable correlation exists between experiencing positive emotions, particularly the emotion of enjoyment, and the active engagement of deep learning strategies among students. This correlation was observed to be closely associated with students adopting a dual approach, including both mastery-oriented and performance-oriented goals within their learning process. The research conducted by Huang et al. [27] reinforces the focus on positive emotions, affirming that emotions like happiness and enjoyment lead to heightened interactions between learners and instructors. The correlation between engagement and emotion data could provide a more nuanced understanding of user experience and by combining these insights, content creators could tailor their strategies to not only capture attention but also evoke positive emotional responses, fostering a more meaningful and impactful connection with their audience.

In line with this context, our focus lies in detecting engagement and emotion during reading tasks, employing a multimodal approach that integrates gaze data and physiological responses from participants. The overview of the design is depicted in Figure 1. In our research, we employed various deep learning models to analyze the multimodal data that has been gathered. This enables us to provide a comprehensive comparison between the efficacy of various models and to comprehend which models are most proficient in handling this complex task. Another objective of our study is to compute the correlation between engagement and emotion during a reading task. Understanding this relationship can provide crucial insights into the dynamic interplay of cognitive and affective processes during reading. By decoding this relationship, we can potentially enrich the design of interactive reading systems and promote improved reading engagement and comprehension strategies.

The key contributions of this study are the following:

- An exhaustive experimental framework for monitoring user engagement and emotional responses during reading tasks, utilizing a non-intrusive multimodal methodology.
- A thorough and comparative evaluation of advanced deep learning models designed for time series prediction to decipher user engagement and emotional states. A user-independent generalized approach and a user-dependent personalized evaluation were implemented for all the models to compare the prediction results.
- Investigating the correlation between user engagement and the emotions elicited during reading tasks.
- A user-oriented application that displays fluctuations in user engagement and emotions through a gauge meter and emojis. This application also provides customized alert messages that assist users in understanding their engagement and emotional shifts.

The user application was designed with a user-centric approach, aiming to visually represent fluctuations in both user engagement and emotions employing a gauge meter, which likely serves as a graphical indicator, and emojis to convey the emotional context. The emojis could add a layer of personalization and expressiveness to the communication of emotions and users may find it more engaging and relatable to see emojis that mirror their feelings, fostering a stronger connection with the application, thus leading to enhanced engagement. The application not only monitors and displays user engagement and emotions, but also actively supports users by providing relevant information and insights to help them navigate and interpret their own emotional and engagement dynamics.

In order, the paper is structured as follows: Section II offers an explanation of the technical background and related work in the domain of detecting engagement and emotions through sensor-based approaches. Section III explains about the user study and data collection methods. Section IV details the techniques and methods used to detect the engagement and emotions in reading, including eye-tracking and physiological data analysis. Section V summarizes the classification performances and the results achieved. Section VI looks at the studies research questions, challenges, and constraints. Lastly, Section VII concludes the paper.

II. BACKGROUND AND RELATED WORK

Engagement and emotion are fundamental aspects that shape the reading experience, reflecting both cognitive investment and affective responses to the content. Both are interconnected and multifaceted aspects of the reading experience. Recently, the fusion of technology with human cognition and emotions has opened new approaches for understanding and enhancing the reading process. This task necessitates a multidisciplinary approach that melds theoretical insights, methodological advancements, and cutting-edge technologies. The following related work section offers

a comprehensive survey of key concepts, techniques, and applications that underline the current study.

A. ENGAGEMENT DETECTION IN READING

Engagement detection in reading is about creating an adaptive, responsive, and student-centric learning environment. Understanding this aspect can have applications in various domains, including education, where tailored strategies can be developed to enhance reading comprehension and enjoyment. Conati et al. conducted an eye-tracking study to analyze how students pay attention to adaptive hints within educational games and their research reveals insights into the effectiveness of the hints in engaging students and guiding their learning, demonstrating that the design and timing of adaptive hints can have a significant impact on attention and educational outcomes [28]. Jacob et al. explored the use of a physiological sensing wristband to detect reader interest while reading newspaper articles and developed a model to assess readers' interest levels, providing insights into personalized content delivery and enhancing user experience in reading digital newspapers [29]. Wang et al. reviewed multi-sensor eye tracking system in the context of capturing student attention in learning [30]. Ishimaru et al. found that the variations in pupil diameter and nose temperature have high correlation with the cognitive states of students, like interest in learning materials in Physics [31]. These studies [32], [33], [34], [35] were centered on the use of sensors and wearables to monitor students' physiological and behavioral patterns, as well as eye-tracking devices to gauge student attention and comprehend engagement in the learning process.

B. EMOTION DETECTION IN READING

Understanding students' emotions can help educators and technology designers create more engaging and personalized learning experiences [36]. D'Mello and Graesser explored the fluctuations of affective states during complex learning, finding that these emotional dynamics play a significant role in the learning process [6]. Schmidt et al. introduced a novel publicly available dataset of multimodal sensor data on stress and affect detection, collected from wearable devices during various activities, enabling researchers to create more robust models for understanding human emotional states [37]. Their findings emphasize the potential applications of this dataset for the development of new algorithms and methods for real-time, continuous stress, and emotion recognition. Calvo, R.A. and D'Mello, S. provide an interdisciplinary review of affect detection, encompassing various models, methods, and applications, highlighting the growing importance and challenges in the field of affective computing [38]. The authors identify a wide range of applications and highlight the existing challenges and limitations in effect detection, such as the complexity of human emotions, the diversity of potential input data (e.g., facial expressions, physiological signals), and the difficulties in accurately modeling and interpreting this data. Arroyo et al. [39] explored the use of emotion sensors in an educational context, specifically focusing on how these sensors can be applied to detect student emotions in a school environment. The findings highlight the potential of using emotion sensors to enhance learning experiences by adapting educational content based on students' emotional responses, although challenges in practical implementation and interpretation of the sensor data were also noted.

C. MULTIMODAL DATA ANALYSIS AND FUSION TECHNIQUES

The multimodal fusion of sensors for detecting cognitive and affective states in reading tasks offers a rich, nuanced understanding of readers' interactions with text [40]. Multimodal affective analysis [41], [42] concentrates on strategies for multimodal integration, categorizing them into early fusion (feature-level), late fusion (decision-level), model-level fusion, and hybrid-level fusion. Nonetheless, these analyses can also diversify based on the combinations of various modalities. Delving deeper, it could be recognized that this approach incorporates a diverse range of data types. From eye-tracking data, which reveals the reader's focus and attention shifts, to physiological measures that offer insights into the underlying emotional and cognitive states, and to behavioral analytics that capture explicit actions and responses, the toolkit for understanding reading experiences becomes robust and multifaceted. However, with such richness in data also comes the inherent challenge of data integration and the ethical implications of collecting and using such detailed personal information. Koelstra et al. in their study, made a significant contribution to this field by introducing a specially curated dataset tailored for the analysis of human affective states. Their research, as detailed in [43], emphasized the added value of integrating physiological signals into the analysis matrix. Their results provided a promising picture where understanding emotions through such measures could dramatically enhance the quality and efficacy of reading and learning interventions. D'Mello et al. offers a comprehensive examination of multimodal affect detection, including techniques for fusing physiological and behavioral data, and the insights obtained could be applied to develop advanced reading engagement and emotion recognition systems [44]. Chen and Parameshachari utilized a multimodal fusion of video, text and physiological signals to construct a model for detecting the learner engagement and the findings stated that the learning engagement evaluation model could accurately assess learning engagement [45]. In their study, Tzirakis et al. [46] utilized a CNN for audio feature extraction and a deep residual network for visual data. These extracted features were then combined and processed through a 2-layer LSTM to estimate Arousal-Valence values. Hao et al. [47] introduced a combined visual-audio emotion detection system that utilized multitask and blending learning techniques with diverse features. This approach employed SVM classifiers and CNNs for both handcrafted and deep learning-based visual-audio features,

resulting in four sub-models. These were then integrated using a blending ensemble algorithm to predict the overall emotion. This method attained average accuracies of 81.36% (speaker-independent) and 78.42% (speaker-dependent) on the eNTERFACE dataset [48].

The advancement and improvement of wearable technologies have amplified interest in automated affective analysis using multiple physiological modalities [49]. Yet, given the intricacy of emotions and considerable variance in individual physiological reactions [50], achieving reliable prediction results with EEG-based or ECG-based emotion detection remains challenging. In studies of multi-physiological modality fusion for affective analysis, in addition to EEG and ECG, other types of physiological signals (e.g., ECG, EOG, BVP, GSR, and EMG) are jointly combined to interpret emotional states [51]. The visual modality (facial expression, voice, gesture, posture, etc.) may also be integrated with multimodal physiological signals for visual-physiological affective analysis [52]. Verma and Tiwary [53] employed a comprehensive approach to emotion estimation by leveraging a variety of physiological data like ECG, GSR, ST, BVP, RESP, EMG, and EOG selected from DEAP. In their methodology, they employed the Discrete Wavelet Transform (DWT) [54] to extract multi-resolution features from this data. These features were then instrumental in estimating emotions across three dimensions: valence, arousal, and dominance.

D. DEEP LEARNING FOR COGNITIVE AND AFFECTIVE COMPUTING

Deep learning has emerged as a pioneering force in the advancement of cognitive and affective computing, reshaping how machines interpret, process, and respond to human emotions and cognitive states. The intricate nature of human emotions and thoughts, including facial expressions, voice modulations, and physiological signals, are adeptly interpreted by deep learning architectures. Particularly, Convolutional Neural Networks (CNNs) [55] and Recurrent Neural Networks (RNNs) [56], along with their advanced variants, have proven to be pivotal in decoding the intricate temporal and spatial interdependencies inherent to emotional and cognitive data.

Martinez et al. implemented end-to-end deep learning for affect recognition using physiological signals [57]. Their approach integrated denoising autoencoders, CNNs, preference learning, and automated feature selection, applied to the Maze-ball dataset. This dataset features BVP and EDA data gathered from individuals playing games, complemented with associated questionnaires. Qiu et al. [58] introduced the Correlated Attention Networks for emotion detection, incorporating bidirectional Gated Recurrent Units (GRUs), a Canonical Correlation Layer, a Signal Fusion Layer, an Attention Layer, and a Classification Layer. They evaluated this structure on three datasets, namely SEED, SEED IV, and DEAP. Badshah et al. [59] introduced a comprehensive deep learning architecture employing an end-to-end deep CNN. This architecture featured three convolutional layers and three fully connected layers designed to interpret the patterns detected by the convolutional layers. Their methodology focused on extracting meaningful features from spectrogram images, which are visual representations of the spectrum of frequencies in a sound signal as they vary with time. Their model was designed to predict seven distinct emotions based on these features. Zhang et al. [60] crafted a design using the AlexNet Deep Convolutional Neural Network (DCNN) pre-trained on ImageNet. To further enhance their model's capability, they integrated a strategy called discriminant temporal pyramid matching (DTPM) which aids in crafting a more global representation of utterance-level features, ensuring the model captures the broader context and nuances of the input data.

Dziezyc et al. [61] evaluated ten distinct deep learning models for affect recognition across four datasets. Their findings indicated that the effectiveness of a model is influenced by the strength of the physiological reactions triggered by emotional stimuli. Additionally, their study suggested that CNN-based structures might be better suited for affect recognition from physiological sensors compared to LSTM-based architectures. Santamaria-Granados et al. [62] presented a study where they employ Deep Convolutional Neural Networks (DCNN) to detect emotions from physiological signals using the AMIGOS [63] dataset. Their findings underline the potential of using deep learning techniques, particularly DCNNs, for emotion recognition tasks, offering insights into the intersection of affective computing and advanced neural network models. Watanabe et al. [64] conducted a study to evaluate the engagement levels of students during online meetings. Utilizing the MobileNetV2, a sophisticated deep learning model, they were able to effectively determine student engagement and their results were promising, with the model achieving an impressive F1-score of 89.5% when tested using the leave-one-participant-out cross-validation method.

E. USER APPLICATIONS AND VISUAL AIDS

User applications and visualization tools dedicated to understanding and illustrating cognitive and emotional states have gained significant attention recently, reflecting the rise in the importance of human-centered design and the increasing sophistication of affective computing technologies. These tools typically harness data from various sources, such as physiological sensors, facial recognition software, and user feedback, to derive insights into a user's cognitive processes and emotional responses. Visualization aids, in particular, transform these complex data streams into intuitive visual representations, making it easier for developers, researchers, and designers to understand and address users' needs, preferences, and emotional triggers. Such advancements are paving the way for more personalized and empathetic digital interactions, bridging the gap between technology and human emotion.

TABLE 1. The self report survey used for the experimental design.

No	Question	Type	Scale
1.	Rate the level of engagement while reading the article.	Engagement	1-5 (1: lowest and 5: highest)
2.	Rate the kind of emotion (positive or negative) experience while reading the article.	Valence	1-5 (1: lowest and 5: highest)
3.	Rate the intensity of the emotion experienced while reading the article	Arousal.	1-5 (1: lowest and 5: highest)

Shaikh et al. [65] introduced an innovative tool designed to assist social science experts in deciphering emotional patterns triggered by significant cultural stressors. This tool enhances the comprehension of the shifts in emotional reactions resulting from these extensive cultural challenges. Watanabe et al. developed an application named EnGauge [64] that visualizes the engagement levels of users during online meetings using a gauge interface. Additionally, a visualization heatmap using Grad-CAM [66] displays the important features inside the input facial image. Selvaraju et al. [66] introduced a new class-discriminative localization method known as Gradient-weighted Class Activation Mapping (Grad-CAM). This technique enhances the transparency of any CNN-based models by offering visual clarifications. Moreover, by merging their Grad-CAM localizations with pre-existing high-resolution visual representations, they were able to generate high-definition class-discriminative Guided Grad-CAM visualizations. Abdul et al. [67] created COGAM to regulate cognitive load in interpretable AI through visual segments, moving beyond merely evaluating the explanations' simplicity and accuracy.

III. USER STUDY AND DATA COLLECTION

The data collection procedure was aimed to capture the participants' subjective responses concerning their engagement, valence, and arousal levels as they performed a reading task. The subjective ratings gathered from the participants were used as the ground truth labels for the data analysis.

A. PARTICIPANTS

The study involved recruiting 20 university students pursuing their Master's degree, comprising 11 male and 9 female participants. Their ages ranged from 22 to 28 years, with an average age of 25. All participants joined the experiment after providing informed consent, and they had the freedom to withdraw from the study at any point if they chose to do so. The detailed information about data consent, sensor usage, and the experiment sessions were communicated to all participants before their participation in the study. The data from 3 participants were discarded due to errors in the data collection, which would be discussed later in the results section.

B. EXPERIMENTAL DESIGN

The experiment involved a carefully crafted comprehension-based reading task, designed to explore the varying levels of engagement and emotional responses evoked in the participants. To achieve this, we curated a diverse collection of 14 different documents, each strategically chosen to elicit distinct emotional experiences and engagement levels. Figure 2 illustrates the experimental setup, where a participant is reading text from an eye-tracker mounted screen and the wristband, recording the physiological data.

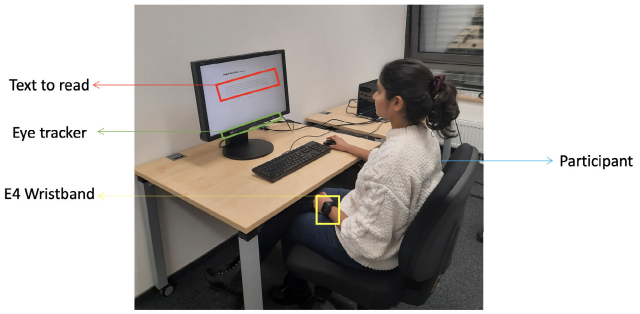


FIGURE 2. The experiment environment.

During the experiment, the participants were presented with each document individually. After reading each document, they were asked to provide subjective ratings for their level of engagement, valence, and arousal using a scale ranging from 1 to 5. On this scale, a rating of 1 indicated the lowest level of engagement, valence, or arousal, while a rating of 5 signified the highest level. Table 1 shows the self-report survey questionnaires and the response scale.

To create a realistic e-learning environment, we developed a web-based document reading platform using technologies such as Next.js, React.js, and Express.js. This digital platform allowed us to create a dynamic and interactive learning setting, ensuring a seamless user experience for the participants. During the experiment, we utilized a Tobii 4C eye tracker with a pro license, which was mounted to a display monitor, to capture the participants' gaze patterns as they read the documents. Additionally, we employed the Empatica E4 wristband to record their physiological responses throughout the reading task. The experiment was divided into two separate sessions, with the participants granted a 10-minute break after reading the first seven documents. The break was introduced to mitigate potential fatigue and ensure that the participants remained alert and engaged throughout the entire duration of the experiment.

IV. METHODOLOGY

Within our system, we implemented a comprehensive methodology that incorporated eye gaze tracking and physiological signal recording during reading sessions in order to assess readers' affective states, including their levels of engagement, arousal, and valence. To accomplish this,

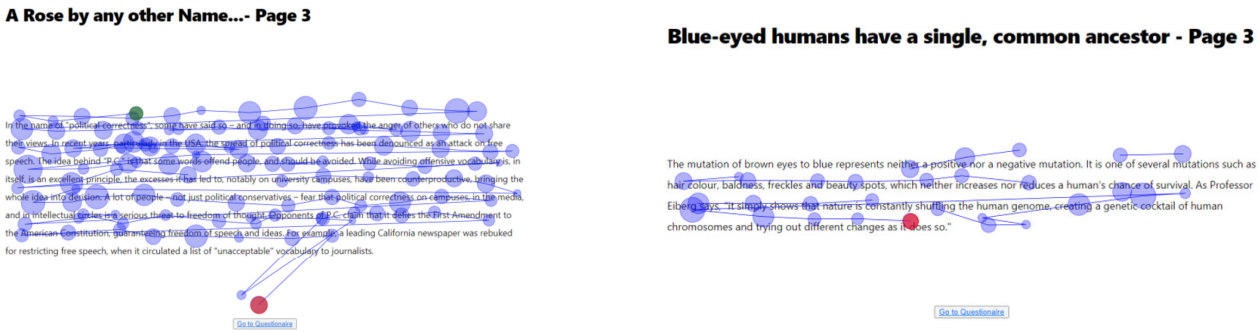


FIGURE 3. Scanpaths of gaze data for highly engaging document (left) vs less engaging document (right) as rated by a participant.

we utilized the Tobii 4C pro remote eye-tracker, capturing gaze coordinates and pupil diameters at a robust sampling rate of 90 Hz. Moreover, we acquired a diverse array of physiological data through the utilization of the E4 wristband sensor, which included 3-axis Acceleration (ACC) at 32Hz, Blood Volume Pulse (BVP) at 64Hz, Electrodermal Activity (EDA) at 4Hz, Skin Temperature (TEMP) at 4Hz, and Heart Rate (HR) at 1Hz. The engagement, valence, and arousal of the participants were detected using both the gaze data and E4 data individually and in combination.

A. DATA PRE-PROCESSING

The gaze data consisted of raw eye attributes like the left and right eye coordinates, along with pupil diameter, all sampled at 90 Hz. To prepare the raw data for analysis, preprocessing steps were taken to remove outliers and noise and a sliding window approach was applied to segment the data into 30-second intervals with a 50% overlap. Figure 3 depicts the fixations and saccades of a participant who rated different documents as highly engaging and least engaging. Similarly, the physiological signals, including EDA, BVP, TEMP, HR, and 3-axis ACC data, were recorded using Empatica E4, with sampling frequencies of 4 Hz, 64 Hz, 4 Hz, 1 Hz, and 32 Hz, respectively. The gaze and physiological signals were synchronized based on the timestamps and preprocessing was carried out on the physiological signals as well, involving noise removal and the use of a sliding window with a length of 30 seconds and a 50% overlap to segment the data for further analysis.

B. MODEL ARCHITECTURE

Deep neural network architectures possess the capability to automatically extract essential information from raw data, thereby reducing the complexity of data processing and minimizing the labor-intensive efforts required for manual feature engineering. The streamlined process of obtaining relevant data boosts the likelihood of acquiring useful insights, which might be elusive to experts in the field. The inherent layer-based structure of neural networks facilitates seamless integration of these layers, enabling the creation of diverse and sophisticated deep learning architectures. In our study, we employed five distinct deep learning architectures to detect and analyze the participants' engagement, arousal, and valence levels. The collected data was analyzed using various models, including Transformers, FCN, Encoder, ResNet and Inception.

The Transformer network architecture, originally designed for natural language processing tasks, has been successfully adapted for time series classification tasks, including those involving gaze and physiological data. In this context, the Transformer network is employed to capture temporal dependencies and relationships in the sequential data. The model's input comprises both individual gaze and physiological data and concatenated versions of both, where each time step is represented as a continuous vector. Positional encoding is incorporated to preserve temporal order information. The core component of the transformer network is the self-attention mechanism, allowing the model to attend to different time steps and capture long-range dependencies in the gaze and physiological signals. The outputs from the self-attention layers are fed through fully connected feed-forward neural networks, enabling the extraction of complex patterns and features relevant to the time series classification task. The architecture's ability to automatically learn representations from the combined gaze and physiological data proves advantageous in discerning subtle temporal patterns and emotional responses, facilitating more accurate and insightful time series classification for applications such as affective computing and human-computer interaction.

The FCN model is composed of three convolutional blocks for each signal, which are subsequently followed by a global average pooling layer. The outputs from these branches are then concatenated and passed through one or more fully connected dense layers. In each convolutional block, a set of filters is applied to the input time series, and the results are passed through a non-linear activation function, such as the ReLU function. The output from the convolutional layers is further processed by the fully connected layers to finally perform the classification task.

The Encoder shares similarities with the FCN, but it incorporates an extra attention layer positioned between the final convolutional blocks and the fully connected dense layers. The attention layer enables the model to focus on important regions or features of the input sequence,

selectively attending to certain elements while suppressing others. By incorporating attention, the Encoder can learn to assign varying levels of importance to different parts of the data, enhancing its ability to recognize salient patterns and features.

The ResNet consists of multiple residual blocks, which facilitate the flow of gradients during training by introducing shortcut connections. The residual blocks can effectively capture complex temporal patterns and dependencies in time series data and enable training by reducing the vanishing gradient problems. Similar to ResNet, the Inception network employs residual blocks, but additionally incorporating inception blocks. These blocks consist of several parallel convolutional layers with various kernel sizes, allowing the network to capture patterns at multiple scales. This design enhances the network's capability to recognize complex features and spatial dependencies in the data, making it particularly effective in handling complex and multiscale patterns often present in real-world datasets.

C. CLASSIFICATION

The deep neural networks were trained to predict the engagement, arousal, and valence responses as reported by the participants. In the case of engagement, participants were asked to rate their experience on a scale from 1 to 5. These ratings were then grouped into two distinct categories: *Low* and *High*. Ratings from 1 to 3 were considered to denote *Low* engagement, while ratings of 4 and 5 were interpreted as *High* engagement. The same approach was followed for predicting the arousal and valence responses. In addition, the arousal and valence responses provided by the participants were grouped into four quadrants, each associated with a distinct emotional state. The quadrants are as follows:

- High Arousal, High Valence (HAHV): This quadrant represents strong, positive emotions such as excitement or joy.
- High Arousal, Low Valence (HALV): This quadrant is indicative of intense, negative emotions such as anger or fear.
- Low Arousal, Low Valence (LALV): This quadrant corresponds to weak, negative emotions such as sadness or boredom.
- Low Arousal, High Valence (LAHV): This quadrant signifies mild, positive emotions like calmness or satisfaction.

D. EVALUATION PROTOCOL

The process of partitioning the data into training and testing sets involved the application of two distinct techniques: user-independent and user-dependent approach. In the user-independent or leave-one-participant-out (LOPO) technique, one participant's data was kept as the test set, while the data from all other participants were utilized for training the model. This process was repeated iteratively for each participant, and the resulting accuracies were averaged to determine the overall model performance. By testing the model on multiple test sets, each consisting of data from different participants, a more robust assessment of the model's generalization to unseen data was achieved. The averaged accuracies across these iterations provided a comprehensive performance evaluation, accounting for individual variations among participants. This approach ensured a more reliable and comprehensive measure of the model's ability to perform well on new and diverse data samples.

For the user-dependent or the person-specific approach, the technique involved a distinct data splitting approach, where data from one participant, with each document read was used as a test set, while all other documents were used for training iteratively and the resulting accuracies were averaged. The same process was repeated for the remaining participants, and the average accuracy was computed across all participants. This strategy considers individual differences in engagement and emotional responses, providing a more personalized approach to detect and analyze individual interests. By considering each person's unique characteristics, this technique enhances the accuracy and relevance of engagement and emotion detection, making it a valuable and tailored method for personalized analysis.

V. RESULTS

For the study, we utilized deep learning models to make predictions on three distinct variables: engagement, valence, and arousal. Two different training methods were employed for these models: user-independent and user-dependent techniques. The prediction targets for engagement, valence, and arousal were categorized through a binary classification task. We derived the 'ground truth', for these tasks, from participant responses to our survey. Moreover, to provide a more comprehensive understanding of emotions, we established a four-class classification structure using valence and arousal responses. This structure segmented the response values into four quadrants, each representing a spectrum of emotions.

A. MODEL PERFORMANCE USING USER-INDEPENDENT APPROACH

After the models were trained, their performance was evaluated in terms of accuracy and F1-score. Table 2 in the report presents these evaluation results for all models that were trained using the user-independent approach. The evaluation metrics were computed in two scenarios: one where each modality (i.e., data type) was considered individually, and another where the modalities were combined. This comparison aimed to assess whether the combination of modalities would yield better predictive performance than using individual modalities.

The Transformer model was found to provide the best performance across all three categories in terms of accuracy and F1-score, compared to the other implemented models. For engagement prediction, specifically, the combination of features derived from eye-tracking data and wristband

TABLE 2. Summary of evaluation metrics using user-independent approach (LOPO).

Models	Sensor	Engagement		Arousal		Valence	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
ResNet	Eye-Tracker	71.25	70.34	67.19	65.21	65.41	64.18
	Wristband	65.72	60.21	58.30	56.10	60.11	59.26
	Combined	74.49	72.16	68.10	66.90	69.93	69.19
Inception	Eye-Tracker	77.56	76.87	65.75	65.15	62.15	59.75
	Wristband	69.12	65.22	56.32	54.22	55.66	54.18
	Combined	78.30	75.06	67.27	63.25	62.75	60.52
FCN	Eye-Tracker	70.21	70.10	66.15	61.65	64.20	63.90
	Wristband	70.90	67.53	55.68	50.35	54.47	53.88
	Combined	72.85	70.18	65.87	62.07	66.28	66.03
Encoder	Eye-Tracker	68.65	64.05	60.84	60.24	61.95	61.42
	Wristband	62.10	62.05	53.26	52.55	54.35	51.38
	Combined	68.90	67.45	60.55	60.10	60.25	59.86
Transformer	Eye-Tracker	79.66	76.11	70.44	68.42	72.82	72.27
	Wristband	71.90	65.29	62.87	61.91	63.66	62.79
	Combined	**80.38**	**78.73**	**71.28**	**71.20**	**73.98**	**71.32**

data was particularly effective. The Transformer model trained on this combined dataset yielded an accuracy of 80.38% and an F1-score of 78.73. These metrics indicate a strong performance by the Transformer model in predicting engagement, surpassing the performance of all other models tested in this study. Similarly, for predicting valence and arousal, the Transformer model performed the best, achieving an accuracy of 71.28% and 73.98% respectively. This performance, like that for engagement prediction, was higher than any of the other models implemented in the study.

It was observed that not only the Transformer model, but also the Inception and ResNet models, which are both well-known deep learning architectures, delivered reasonably good performances. Interestingly, our results show that when the sensor signals were combined, rather than used individually, there was a boost in performance for all the three target variables. This demonstrates the potential of a multimodal approach in improving the accuracy of predictions in complex tasks such as emotion and engagement detection. However, while analyzing individual sensor data, we observed some variability in model performance. The models demonstrated a lower performance when solely utilizing data from the wristband, which could be due to the lower sampling frequencies of the E4 wristband. The Empatica E4 captures data at lower frequencies compared to other data collection methods, such as eye-tracking devices. This could potentially lead to a less rich dataset, missing some finer details that might be important for accurately predicting engagement, arousal, and valence. The results of this classification can be seen in Figure 4a, which displays a confusion matrix visualizing the performance of our binary classification model for engagement prediction using the user-independent approach. Figure 4b displays a confusion matrix for the quadrant-based four-class classification, providing insights into how accurately our model was able to predict the emotional states of the participants based on their arousal and valence values using the user-independent approach.

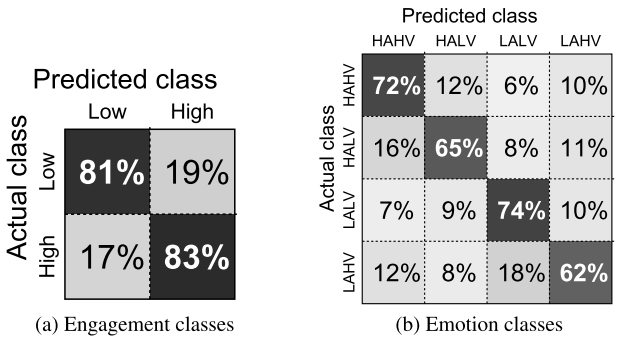


FIGURE 4. Confusion matrices for engagement classification (low and high) and emotion classification (HAHV: High Arousal High Valence, HALV: High Arousal Low Valence, LALV: Low Arousal Low Valence, and LAHV: Low Arousal High Valence) using user-independent approach.

B. MODEL PERFORMANCE USING USER-DEPENDENT APPROACH

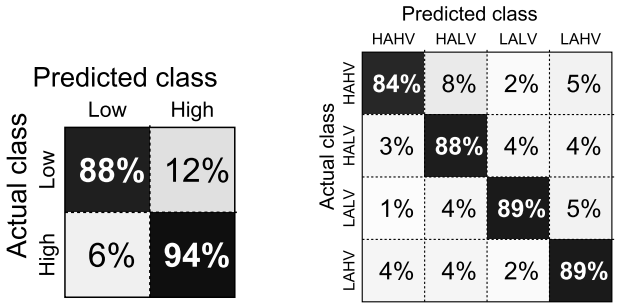
Our research also focuses on the idea that engagement and emotions are deeply individualized experiences, subject to variation across different individuals even while performing identical tasks. Each individual possesses unique characteristics and patterns of emotional and cognitive response, contributing to a distinctive profile of engagement and emotional responses. Consequently, detecting and predicting these experiences using a generalized model may not yield optimal results due to the lack of personalization. To address this matter and tailor our predictive models to the

TABLE 3. Summary of evaluation metrics using user-dependent approach.

Models	Sensor	Engagement		Arousal		Valence	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
ResNet	Eye-Tracker	89.25	85.22	88.80	86.17	87.78	87.17
	Wristband	85.67	81.28	84.26	82.11	83.16	81.16
	Combined	**93.56**	**91.78**	**90.62**	**89.85**	**88.70**	**88.32**
Inception	Eye-Tracker	87.30	86.39	85.10	82.93	82.90	80.91
	Wristband	80.42	78.30	82.26	81.16	79.56	79.03
	Combined	88.26	87.83	87.62	86.56	83.12	82.72
FCN	Eye-Tracker	85.43	84.19	84.51	84.22	84.19	83.93
	Wristband	81.28	80.22	79.96	78.19	81.32	80.22
	Combined	84.75	83.76	84.55	84.50	83.77	83.54
Encoder	Eye-Tracker	86.66	86.10	86.29	85.09	83.46	81.18
	Wristband	80.05	79.82	82.04	80.36	78.88	78.10
	Combined	85.52	84.91	84.78	82.34	82.16	81.32
Transformer	Eye-Tracker	88.10	86.37	88.10	85.21	86.50	86.22
	Wristband	83.54	80.21	82.78	80.64	81.41	80.71
	Combined	90.81	89.76	89.35	88.83	86.93	86.50

specificities of individual experiences, we have additionally implemented a user-dependent approach. This approach emphasizes personalization, incorporating user-specific traits and patterns into the predictive models to create a unique model for each individual. It stands in contrast to the user-independent approach, which typically trains models on aggregate data from multiple users without accounting for individual differences.

Our results indicated that this personalized, user-dependent approach was indeed effective, which can be observed from Table 3. Across all the deep learning models we employed, the predictions for engagement, valence, and arousal were consistently higher using the user-dependent approach compared to the user-independent approach. The ResNet model with the combined features from the two modalities achieved the highest accuracy of 93.56%, 90.62% and 88.70% for engagement, arousal, and valence predictions respectively. These findings affirm the importance of personalization in emotion and engagement detection, supporting the notion that tailored, user-specific models can significantly improve prediction performance. This insight has implications not only for the development of more effective and responsive reading systems, but also for a wider range of applications where understanding and predicting user engagement and emotional responses are critical. Our findings suggest that embracing individual variability and focusing on personalization may be key strategies in the advancement of emotion and engagement detection technology. The classification outcomes are illustrated in Figure 5a, using a confusion matrix that visually represents our binary classification model predicting engagement through the user-dependent approach. Additionally, Figure 4b shows the confusion matrix for the quadrant-based four-class classification, predicting participants' emotional states based on arousal and valence values, employing the user-dependent approach.



(a) Engagement classes (b) Emotion classes

FIGURE 5. Confusion matrices for engagement classification (low and high) and emotion classification (HAHV: High Arousal High Valence, HALV: High Arousal Low Valence, LALV: Low Arousal Low Valence, and LAHV: Low Arousal High Valence) using user-dependent approach.

C. INTERPRETING DOCUMENT AND PARTICIPANT ENGAGEMENT PATTERNS

The heatmap shown in Figure 6 provides a visualization of the average *Engagement* ratings given by different participants (pId) for various documents (docId). The color of each cell in the heatmap reflects the average *Engagement* rating given by a specific participant to a specific document. The color map used determines the relationship between the engagement score and the color in the heatmap. The darker color represents less engagement, and a lighter color represents high engagement. By observing the color patterns in the heatmap, trends can be identified and could gain insights. For example, if certain documents consistently receive high engagement scores across all participants, these documents might be particularly interesting or well-written. Similarly,

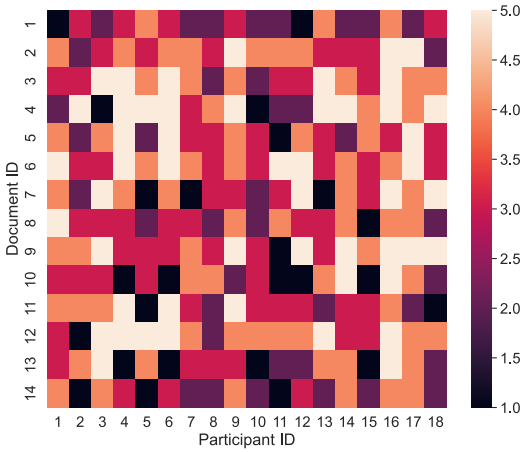


FIGURE 6. The heatmap of average engagement ratings provided by each participant to the documents.

if some participants show high engagement levels with all documents, they may be particularly enthusiastic or interested in the study's subject matter.

Taking a closer look at the heatmap, it can be noticed that Document 1 (D1) received predominantly darker colors across its row. This implies that this document was generally less engaging to most participants. In other words, when averaged across all participants, D1 received lower engagement ratings compared to other documents. This might indicate that the content, presentation, or subject matter of D1 was less appealing or stimulating to the study participants. In contrast, D12 is predominantly associated with lighter colors, indicating higher engagement ratings from a majority of the participants. It seems that D12 was particularly successful in capturing the attention and interest of the participants, suggesting that its content was engaging and stimulating to the participants.

In addition to insights about document engagement, the heatmap also provides information about participant behavior. The prevalence of lighter shades in the documents suggests that Participant 16 (P16) provided higher engagement ratings to almost all the documents. This suggests that this participant was generally more engaged with the documents, or perhaps more generous with their ratings, compared to other participants. On the other hand, P11 is associated with darker shades across most of their ratings, indicating that they found a majority of the documents less engaging. This could be because the participant was more critical or less engaged in general, or possibly the documents did not align well with their interests.

VI. DISCUSSION

In this section, we delve into the key contributions, outcomes, and insights that have been achieved through this study. This includes both the theoretical implications of the research and the practical applications that might result from it. This section considers the uniqueness of the data set, the novel approach to analysis, and the relevance of the research question to ongoing discussions in the field. This section also provides insights into the study by critically examining its challenges and limitations.

A. EXPERIMENTAL PROTOCOL FOR ENGAGEMENT AND EMOTION DETECTION DURING READING TASK

In many prior studies, the focus has been on detecting engagement in reading or e-learning environments, or on the detection of emotion during tasks such as watching videos. In this study, we sought to bring these two strands together, integrating the responses of both engagement and emotions, specifically arousal and valence, within a reading task. The challenge lies in accurately estimating emotions during reading tasks, which can be more complex than interpreting emotional responses while watching videos.

To address this, our experimental design included not just emotional responses, but also engagement responses in the survey. Furthermore, we employed unobtrusive data collection techniques to ensure the most accurate results possible. We utilized a stationary eye-tracker mounted on a display screen to record the eye movements of participants, a method that provides vital clues about engagement and focus levels. In addition, an Empatica E4 wristband was used to record the physiological responses of the participants synchronously, providing real-time data about the emotional state and engagement level of the readers. This combination of tracking engagement and emotional responses provides a more holistic view of the reading experience, potentially offering insights on how to optimize learning environments or understand more about the cognitive and emotional processes involved in reading.

In order to elicit diverse engagement and emotional responses from our participants, we carried out a meticulous selection of documents via sentiment analysis. This methodology was employed to ensure that the reading material could evoke a broad spectrum of emotional responses, mainly happiness, anger, boredom, and calmness. The documents chosen included a variety of genres and types, including scientific articles, literary works, fiction, short stories, and poems. This diverse range was designed to appeal to different reader preferences and elicit a wider range of emotional and engagement responses, thereby providing more comprehensive data for our study.

To minimize potential sources of data noise and outliers, we made sure to display the documents one paragraph at a time in a centered layout. This specific presentation format was used to ensure the recording of precise eye movements and having only one paragraph displayed at a time helped keep the reader's focus relatively steady and centralized, reducing the likelihood of erratic eye movements and thereby providing cleaner, more reliable data. This setup not only allowed us to capture detailed eye tracking and physiological data but also helped ensure that the range of emotional and engagement responses we recorded were reflective of genuine reactions to diverse reading material. This combination of detail-oriented protocol and diversity

of stimuli contributes to the robustness and validity of our findings.

B. USER-INDEPENDENT AND USER-DEPENDENT PREDICTION USING DEEP LEARNING MODELS

The raw features after pre-processing from both the sensing modalities; eye-tracker and wristband were utilized to train different state-of the art times series classification models to detect the elicited engagement, valence, and arousal responses. Our models performed exceptionally well in predicting engagement in comparison to predicting valence and arousal. We used these sensing modalities both individually and combined to predict the target variables. Intriguingly, we discovered that employing a combination of sensors offered superior performance, as reflected in higher accuracy and F1-scores. Two distinct approaches were adopted for our study. The first was a user-independent approach, which made use of Leave-One-Participant-Out (LOPO) cross-validation, and the second was a user-dependent approach. This allowed us to draw comparisons between the outcomes of a generalized predictive model and a personalized model.

In terms of performance, the Transformer network stood out from other models, particularly when utilizing a combination of features to predict engagement, valence, and arousal under a user-independent approach. Although the predictions for engagement were robust, we noticed that the predictions for valence and arousal still required additional refinements. Interestingly, when the modalities were used individually, gaze features served as stronger predictors for all target variables as compared to the physiological signals recorded during a reading task. This may be due to the wristband recording signals at a lower frequency, which could have contributed to lower performance metrics for the physiological signals.

The implementation of the user-dependent approach yielded a marked increase in performance across all models. More specifically, we observed a significant rise in accuracy across three main metrics: a substantial 13% boost in predicting user engagement, a 19% boost in predicting arousal, and a 15% leap in valence prediction. In terms of overall performance, it was the ResNet model that delivered the best results as per our evaluation metrics for all the target variables, outperforming the other models. Moreover, the introduction of the user-dependent approach contributed to improving predictions of valence and arousal. This served to underscore the potential for adopting personalized models in user studies, as these can significantly enhance the accuracy of predictions. The outcomes suggest a promising direction for future research in user experience design and personalized systems. By tailoring models to individual users, we may improve the predictive capacity of these models and consequently improve user interaction. This could result in more engaging and effective user experiences, which is a critical aspect in fields such as e-learning, digital marketing, and user interface design.

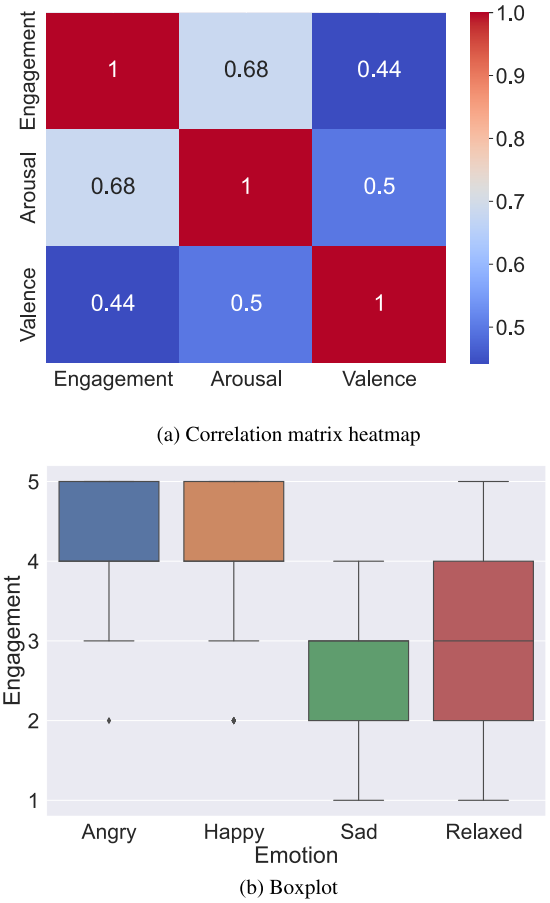


FIGURE 7. The correlation matrix heatmap showing the linear relationship between engagement, arousal, and valence and a boxplot to visualize the distribution of engagement across various emotional categories.

C. ENGAGEMENT AND EMOTION CORRELATION IN READING TASK

One of the main aims for this study was to estimate the correlations between the engagement and emotions experienced by users in a reading task. The participants rated their responses on a scale from 1-5 for engagement, valence, and arousal, with '1' being the least and '5' being the highest.

In Figure 7a, a correlation matrix is represented as a heatmap, offering a visual overview of the linear associations between the pairs of variables - Engagement, Arousal, and Valence. In the case of engagement and arousal, a correlation coefficient of 0.68 is observed, suggesting a relatively strong positive correlation. This implies that an increase in engagement tends to be associated with an increase in arousal, and vice versa. Similarly, the correlation coefficient between engagement and valence is 0.44, indicating a moderate positive correlation. This means that as engagement levels rise, valence also typically tends to increase, and the same holds true in reverse.

Figure 7b represents a boxplot for visualizing and comparing the distributions of *Engagement* values across different *Emotion Categories*. This type of plot enables us to see variations in Engagement levels based on the type of emotion

experienced by participants. From this boxplot, we can infer that emotions categorized as *happiness* or *excitement* (characterized by high arousal and high valence) and *anger* or *fear* (characterized by high arousal and low valence) are generally associated with higher levels of engagement. On the other hand, emotions such as *sadness* or *tiredness*, associated with low arousal and low valence levels, typically correspond to lower levels of engagement. Lastly, when participants reported feeling *relaxed* or *calm*, their engagement levels fell in a moderate range, neither too high nor too low. Therefore, the emotional state of the participant appears to significantly influence their engagement levels, highlighting the need for a thorough consideration of the emotional context when designing and interpreting user engagement studies.

Based on the results drawn from our study's analyses, it becomes apparent that engagement and emotions together constitute a pivotal element in contexts of learning or reading. The interaction of emotions and engagement presents a more layered understanding of the learning process. High arousal emotional states, whether positive or negative, tend to increase engagement levels. Conversely, low arousal states tend to correlate with lower engagement. Thus, our study underscores the importance of a more holistic approach in learning and reading environments. Both emotions and engagement should be considered to optimize learning processes and outcomes. This understanding could be crucial for educators and instructional designers to develop effective strategies that cater to the emotional states of learners, thereby enhancing engagement and overall learning success.

D. USER APPLICATION

Our study introduces an interactive and user-friendly, dashboard designed to provide a visual representation of a reader's engagement level and emotional state while performing a reading task. The dashboard features an engagement gauge that displays the reader's level of engagement based on predicted class probabilities. These predictions were made using the best performing model, utilizing a user-dependent approach. The purpose of adopting a user-dependent approach was to create a more personalized experience for the reader, allowing for the most accurate representation of fluctuations in their engagement levels. The gauge scale on the dashboard spans from 0 to 100, thereby providing a comprehensive numerical range within which the reader's level of engagement can be assessed and visualized. This graphical depiction of engagement levels allows users to evaluate their interaction with the material they have read. In addition to the engagement gauge, the dashboard also includes an emotion emoji. This serves as a visual cue that illustrates the predicted emotional state of the reader. The emotion emoji was determined based on the best-performing model in terms of accuracy, which predicts both valence and arousal - the two dimensions typically used to categorize human emotions.

FIGURE 8. The dashboard notifying the participant to concentrate on text as it detected a fall in engagement level.

FIGURE 9. The dashboard motivating the participant to continue doing the good work since the engagement levels are high over a consistent period.

In addition, the dashboard is equipped with an innovative alert system. This system has been created to assist users in sustaining an optimal engagement level while reading. If a situation arises where a user's engagement level falls below a set threshold, the system promptly generates an alert message, informing the user of the drop in their engagement. Conversely, if a user's engagement level consistently rises above a specific value, indicating a high level of engagement, the system will trigger a motivational message. This feature serves to further boost the user's morale and encourage the continuation of their elevated engagement levels. This interactive platform not only provides users with immediate visual and quantitative feedback on their engagement and emotional states, but also actively aids in the maintenance and improvement of high engagement levels with its alert and motivation system.

Figure 8 displays a screenshot of our user-interface portraying an instance where the application alerts the user to increase their focus on the text, due to a drop in their predicted engagement levels below a predetermined threshold value. Such a feature could be instrumental in promptly drawing the user's attention to their lessening engagement, thus allowing them the opportunity to consciously readjust their focus and re-engage with the task at hand. On the other hand, Figure 9 demonstrates an instance of the dashboard providing the user with a motivational message. This situation arises when the user's engagement levels are consistently surpassing the set threshold, indicating a high level of engagement with the reading task. The motivational message was designed to capitalize on these moments of heightened engagement, reinforcing the user's morale and promoting the activity of their elevated engagement levels.

This tool, we believe, could prove particularly beneficial in the context of e-learning environments. By providing personalized feedback and interactive features that foster high engagement levels, the dashboard can potentially

revolutionize how engagement is approached and cultivated in e-learning environments.

E. LIMITATIONS AND CHALLENGES

While the described study makes significant contributions to the understanding of user engagement and emotional responses during reading tasks, there are certain limitations and challenges that should be considered:

Generalization Across User Groups: The study mentions a user-independent generalized approach, but it may still be challenging to ensure that the findings generalize well across diverse user groups. Users vary widely in terms of preferences, demographics, and individual differences, and the effectiveness of the models may differ for different user segments.

Subjectivity in Emotion Assessment: Assessing emotions can be inherently subjective. Different individuals may interpret and express emotions differently, and relying on self-reported emotions or external indicators like emojis might not capture the full complexity of users' emotional states. Ensuring the reliability and validity of emotional assessments is an ongoing challenge.

Model Robustness and Generalization: The thorough evaluation of deep learning models is mentioned, but the robustness of these models in real-world scenarios and their generalization to different reading contexts may be a limitation. The models' performance in controlled experimental settings might not fully reflect their effectiveness in diverse and dynamic reading environments.

Correlation vs. Causation: Investigating the correlation between user engagement and emotions is valuable, but establishing a causal relationship is complex. While the study may reveal associations, it may not definitively identify whether changes in engagement cause changes in emotions or vice versa.

Real-time Monitoring Challenges: The application's goal of real-time monitoring is ambitious and may face challenges related to latency, real-world distractions, and user interruptions. Achieving true real-time insights might be difficult, and the application's effectiveness may depend on seamless integration into users' reading routines.

Alert Message Effectiveness: While the application provides customized alert messages, the effectiveness of these messages in assisting users to understand their engagement and emotional shifts needs to be validated. User response to alerts may vary, and the impact on behavior or emotional regulation should be carefully assessed.

Long-term User Engagement: The study's focus on short-term reading tasks may limit its insights into long-term user engagement and emotional patterns. Understanding how these dynamics evolve over extended periods could provide a more holistic view of user experiences.

Addressing these limitations and challenges could strengthen the study's impact and contribute to the ongoing discourse on user engagement, emotions, and the application of deep learning in understanding human behavior during reading tasks.

VII. CONCLUSION

This research paper contributes to the understanding and detection of engagement and emotion during reading tasks, using a multimodal approach that combines gaze data and physiological responses from participants. A wide-range range of advanced deep learning models utilized to process and analyze the gathered multimodal data, has facilitated a comprehensive comparison of these models' effectiveness in performing such intricate tasks. A comprehensive experimental framework was designed that leverages a non-intrusive multimodal approach to record the user responses to engagement, valence, and arousal. We implemented both a user-independent approach for general applicability and a user-dependent approach for personalized predictions. This comparative study facilitated the understanding of these models' performance in such complex tasks. The research delved into the correlation between user engagement and the emotions evoked during reading tasks. This understanding provided insights into how emotions and cognitive engagement intertwine during a reading task, helping in developing strategies to enhance user engagement. Finally, the study utilized the prediction results, in the creation of a user-centric application that visually represents fluctuations in user engagement and emotions through a gauge meter and emojis. This application further assists users by providing customized alert messages, enabling users to understand their engagement levels and emotional shifts better.

REFERENCES

[1] S. D'Mello, R. Dale, and A. Graesser, "Disequilibrium in the mind, disharmony in the body," *Cognition Emotion*, vol. 26, no. 2, pp. 362–374, Feb. 2012.

[2] R. Bixler and S. D'Mello, "Automatic gaze-based user-independent detection of mind wandering during computerized reading," *User Model. User-Adapted Interact.*, vol. 26, no. 1, pp. 33–68, Mar. 2016.

[3] N. Jaques, C. Conati, J. M. Harley, and R. Azevedo, "Predicting affect from gaze data during interaction with an intelligent tutoring system," in *Proc. Int. Conf. Intell. Tutoring Syst. (ITS)*, Honolulu, HI, USA. Cham, Switzerland: Springer, Jun. 2014, pp. 29–38.

[4] M.-T. Wang and J. S. Eccles, "School context, achievement motivation, and academic engagement: A longitudinal study of school engagement using a multidimensional perspective," *Learn. Instruct.*, vol. 28, pp. 12–23, Dec. 2013.

[5] J. J. Appleton, S. L. Christenson, D. Kim, and A. L. Reschly, "Measuring cognitive and psychological engagement: Validation of the student engagement instrument," *J. School Psychol.*, vol. 44, no. 5, pp. 427–445, Oct. 2006.

[6] S. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," *Learn. Instruct.*, vol. 22, no. 2, pp. 145–157, Apr. 2012.

[7] S. Ishimaru, S. 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.

[8] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. E. Kaliouby, "AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, May 2016, pp. 3723–3726.

[9] D. McDuff, R. el Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Affectiva-MIT facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected 'in-the-wild,'" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 881–888.

[10] R. W. Picard, "Affective computing: Challenges," *Int. J. Hum.-Comput. Stud.*, vol. 59, nos. 1–2, pp. 55–64, 2003.

[11] M.-Z. Poh, N. C. Swenson, and R. W. Picard, "A wearable sensor for unobtrusive, long-term assessment of electrodermal activity," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 5, pp. 1243–1252, May 2010.

[12] V. Skaramagkas, G. Giannakakis, E. Ktistakis, D. Manousos, I. Karatzanis, N. S. Tachos, E. Tripoliti, K. Marias, D. I. Fotiadis, and M. Tsiknakis, "Review of eye tracking metrics involved in emotional and cognitive processes," *IEEE Rev. Biomed. Eng.*, vol. 16, pp. 260–277, 2023.

[13] 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.

[14] B. W. Miller, "Using reading times and eye-movements to measure cognitive engagement," *Educ. Psycholog.*, vol. 50, no. 1, pp. 31–42, Jan. 2015.

[15] 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.

[16] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[17] R. L. Mandryk and M. S. Atkins, "A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies," *Int. J. Hum.-Comput. Stud.*, vol. 65, no. 4, pp. 329–347, Apr. 2007.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-Resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, Feb. 2017, pp. 1–7.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[20] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, "DeepSD: Generating high resolution climate change projections through single image super-resolution," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 1663–1672.

[21] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[23] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[24] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Apr. 2018.

[25] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[26] E. D. Dela Rosa and A. B. I. Bernardo, "Are two achievement goals better than one? Filipino students' achievement goals, deep learning strategies and affect," *Learn. Individual Differences*, vol. 27, pp. 97–101, Oct. 2013.

[27] J. Yu, C. Huang, Z. Han, T. He, and M. Li, "Investigating the influence of interaction on learning persistence in online settings: Moderation or mediation of academic emotions?" *Int. J. Environ. Res. Public Health*, vol. 17, no. 7, p. 2320, 2020.

[28] C. Conati, N. Jaques, and M. Muir, "Understanding attention to adaptive hints in educational games: An eye-tracking study," *Int. J. Artif. Intell. Educ.*, vol. 23, nos. 1–4, pp. 136–161, Nov. 2013.

[29] 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.*, Oct. 2018, pp. 78–81.

[30] Y. Wang, S. Lu, and D. Harter, "Multi-sensor eye-tracking systems and tools for capturing student attention and understanding engagement in learning: A review," *IEEE Sensors J.*, vol. 21, no. 20, pp. 22402–22413, Oct. 2021.

[31] 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.

[32] A. Apicella, P. Arpaia, M. Frosolone, G. Improta, N. Moccaldi, and A. Pollastro, "EEG-based measurement system for monitoring student engagement in learning 4.0," *Sci. Rep.*, vol. 12, no. 1, p. 5857, Apr. 2022.

[33] Y. Lu, S. Zhang, Z. Zhang, W. Xiao, and S. Yu, "A framework for learning analytics using commodity wearable devices," *Sensors*, vol. 17, no. 6, p. 1382, Jun. 2017.

[34] M. Carroll, M. Ruble, M. Dranias, S. Rebensky, M. Chaparro, J. Chiang, and B. Winslow, "Automatic detection of learner engagement using machine learning and wearable sensors," *J. Behav. Brain Sci.*, vol. 10, no. 3, pp. 165–178, 2020.

[35] J. Oh and H. Kang, "User engagement with smart wearables: Four defining factors and a process model," *Mobile Media Commun.*, vol. 9, no. 2, pp. 314–335, May 2021.

[36] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Automatically recognizing facial indicators of frustration: A learning-centric analysis," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 159–165.

[37] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 400–408.

[38] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[39] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson, "Emotion sensors go to school," in *Artificial Intelligence in Education*. Amsterdam, The Netherlands: Ios Press, 2009, pp. 17–24.

[40] N. Alyuz, E. Okur, U. Genc, S. Aslan, C. Tanriover, and A. A. Esme, "An unobtrusive and multimodal approach for behavioral engagement detection of students," in *Proc. 1st ACM SIGCHI Int. Workshop Multimodal Interact. Educ.*, New York, NY, USA, Nov. 2017, pp. 26–32.

[41] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, and W. Zhang, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Inf. Fusion*, vols. 83–84, pp. 19–52, Jul. 2022.

[42] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. M. Rahaman, and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," *J. Netw. Comput. Appl.*, vol. 149, Jan. 2020, Art. no. 102447.

[43] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis ;Using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[44] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, pp. 1–36, Apr. 2015.

[45] J. Chen and B. D. Parameshachari, "Construction of a learning engagement evaluation model based on multi-modal data fusion," in *Proc. Int. Conf. Knowl. Eng. Commun. Syst. (ICKES)*, Dec. 2022, pp. 1–7.

[46] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[47] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, and P. Xiao, "Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features," *Neurocomputing*, vol. 391, pp. 42–51, May 2020.

[48] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE 05 audio-visual emotion database," in *Proc. 2nd Int. Conf. Data Eng. Workshops (ICDEW)*, Feb. 2006, p. 8.

[49] P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven, "Wearable-based affect Recognition—A review," *Sensors*, vol. 19, no. 19, p. 4079, Sep. 2019.

[50] X. Zhang, J. Liu, J. Shen, S. Li, K. Hou, B. Hu, J. Gao, T. Zhang, and B. Hu, "Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4386–4399, Sep. 2021.
[51] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition," *Inf. Process. Manage.*, vol. 57, no. 3, May 2020, Art. no. 102185.
[52] D. Wu, J. Zhang, and Q. Zhao, "Multimodal fused emotion recognition about expression-EEG interaction and collaboration using deep learning," *IEEE Access*, vol. 8, pp. 133180–133189, 2020.
[53] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage*, vol. 102, pp. 162–172, Nov. 2014.
[54] C. E. Heil and D. F. Walnut, "Continuous and discrete wavelet transforms," *SIAM Rev.*, vol. 31, no. 4, pp. 628–666, Dec. 1989.
[55] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification," *J. Syst. Eng. Electron.*, vol. 28, no. 1, pp. 162–169, Feb. 2017.
[56] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. 5th Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2015, pp. 73–80.
[57] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2017, pp. 985–990.
[58] J.-L. Qiu, X.-Y. Li, and K. Hu, "Correlated attention networks for multimodal emotion recognition," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 2656–2660.
[59] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. Int. Conf. Platform Technol. Service (PlatCon)*, Feb. 2017, pp. 1–5.
[60] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, Jun. 2018.
[61] M. Dzieżyc, 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.
[62] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS)," *IEEE Access*, vol. 7, pp. 57–67, 2019.
[63] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Apr. 2021.
[64] K. Watanabe, T. Sathyanarayana, A. Dengel, and S. Ishimaru, "EnGauge: Engagement gauge of meeting participants estimated by facial expression and deep neural network," *IEEE Access*, vol. 11, pp. 52886–52898, 2023.
[65] S. Shaikh, K. Ravi, T. Gallicano, M. Brunswick, B. Aleshire, O. El-Tayeby, and S. Levens, "EmoVis—An interactive visualization tool to track emotional trends during crisis events," in *Proc. Int. Conf. Appl. Hum. Factors Ergonom.*, Washington, DC, USA, Cham, Switzerland: Springer, 2020, pp. 14–24.
[66] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[67] A. Abdul, C. von der Weth, M. Kankanhalli, and B. Y. Lim, "COGAM: Measuring and moderating cognitive load in machine learning model explanations," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–14.

JAYASANKAR SANTHOSH (Member, IEEE) 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. He is currently pursuing the Ph.D. degree with the German Research Center for Artificial Intelligence (DFKI), Kaiserslautern.

He was a Research Assistant with DFKI, from 2017 to 2019. He has been a Teaching Assistant with the University of Kaiserslautern-Landau, since 2022. He is currently a member of the Immersive Quantified Learning Laboratory (IQL Lab) and the Smart Data and Knowledge Services (SDS) Department, DFKI. He has published papers at IEEE Access, Ubicomp Conference, and IUI Conference. 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).

AKSHAY PALIMAR PAI was born in Bengaluru, India, in 1998. He received the bachelor's degree in computer science from the Ramaiah University of Applied Sciences, India, in 2020, and the master's degree in computer science from the University of Kaiserslautern, in 2023. From October 2021 to September 2022, he was a Project Student and a Research Assistant with the Immersive Quantified Learning Laboratory (IQL Lab) and the Smart Data and Knowledge Services (SDS) Department, DFKI, Kaiserslautern, Germany, focusing on feedback-based intervention in e-learning using deep learning on eye-tracker and physiological sensors. Since 2022, he has shifted his focus of studies to digital farming and completed his internship at Xarvio BASF Digital Farming, where he worked on automating the development process of crop phenology prediction models. He is currently a Junior Engineer with the Smart Data and Knowledge Services Department in German Research Center for Artificial Intelligence (DFKI). His current research interests include the development of a platform to build deep learning models using satellite images and earth observation data for a broad field of applications.

SHOYA ISHIMARU (Member, IEEE) was born in Ehime, Japan, in 1991. He received the B.E. and M.E. degrees in electrical engineering and 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.

He has been a Researcher with the Keio Media Design Research Institute, since 2014. He has been a Senior Researcher with the German Research Center for Artificial Intelligence (DFKI), Germany, since 2019. He served as a Junior Professor with the University of Kaiserslautern-Landau, Germany, from 2021 to 2023. He has been a Project Professor with the Department of Computer Science, Osaka Metropolitan University, since 2023. His research interests include human–computer interaction, machine learning, and cognitive psychology, with the aim of amplifying human intelligence.

Prof. Ishimaru's awards and honors include the Best Presentation Award at the Asian CHI Symposium, in 2020, the Poster Track Honorable Mention at UbiComp/ISWC, in 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 in Japan.