IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 September 2024, accepted 3 December 2024, date of publication 11 December 2024,
date of current version 26 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3515838

RESEARCH ARTICLE

Estimating Self-Confidence in Video-Based Learning Using Eye-Tracking and Deep Neural Networks

ANKUR BHATT[1,2], KO WATANABE[1,2], JAYASANKAR SANTHOSH[1,2], (Member, IEEE), ANDREAS DENGEL[1,2], AND SHOYA ISHIMARU[3], (Member, IEEE)

[1]RPTU Kaiserslautern-Landau, 67663 Kaiserslautern, Germany
[2]German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany
[3]Osaka Metropolitan University, Naka-ku, Sakai, Osaka 599-8531, Japan

Corresponding author: Ankur Bhatt (ankur.bhatt@dfki.de)

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Self-confidence is a crucial trait that significantly influences performance across various life domains, leading to positive outcomes by enabling quick decision-making and prompt action. Estimating self-confidence in video-based learning is essential as it provides personalized feedback, thereby enhancing learners' experiences and confidence levels. This study addresses the challenge of self-confidence estimation by comparing traditional machine-learning techniques with advanced deep-learning models. Our study involved a diverse group of thirteen participants (N=13), each of whom viewed and provided responses to seven distinct videos, generating eye-tracking data that was subsequently analyzed to gain insights into their visual attention and behavior. To assess the collected data, we compare three different algorithms: a Long Short-Term Memory (LSTM), a Support Vector Machine (SVM), and a Random Forest (RF), thereby providing a comprehensive evaluation of the data. The achieved outcomes demonstrated that the LSTM model outperformed conventional hand-crafted feature-based methods, achieving the highest accuracy of 76.9% with Leave-One-Category-Out Cross-Validation (LOCOCV) and 70.3% with Leave-One-Participant-Out Cross-Validation (LOPOCV). Our results underscore the superior performance of the deep-learning model in estimating self-confidence in video-based learning contexts compared to hand-crafted feature-based methods. The outcomes of this research pave the way for more personalized and effective educational interventions, ultimately contributing to improved learning experiences and outcomes.

INDEX TERMS Eye-tracking, learning augmentation, self-confidence estimation.

I. INTRODUCTION

The COVID-19 pandemic has significantly disrupted traditional learning methods [1], highlighting the need for effective online teaching alternatives and approaches. Enhancing existing e-learning platforms [2] is significant to meet academic requirements. This transition increases awareness of censorship and surveillance [3], but on the other hand, it opens up accessibility for anyone to access online lectures. The possibility of students losing self-confidence due to the lack of feedback and the non-verbal behaviors of traditional classes may be challenging, eventually inhibiting students' desire to learn and improve their grades.

Quantified learning has incredible potential in the era of digital education. It allows us to monitor learning behaviors and provide specific feedback to both teachers and learners. Smart sensors in devices like computers, tablets, smartphones [4], chairs [5], and even eyeglasses [6] allow access to students' physical and cognitive states while learning. Physical states include various nonverbal cues, like utterance

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

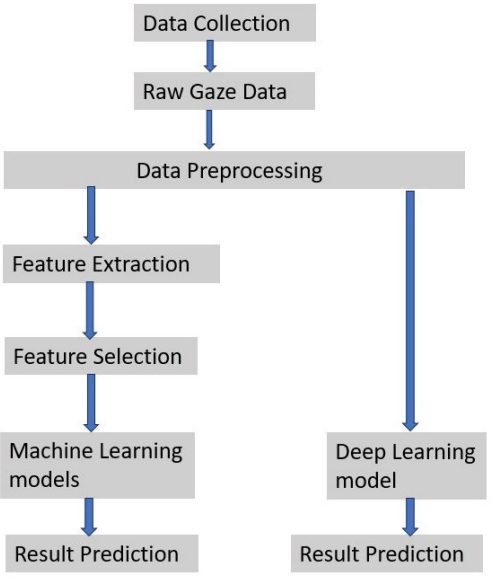


FIGURE 1. An overview of the data collection to analysis.

rate [7], [8], nodding [9], [10], [11], and smiling [12], [13], which provides insights into an individual's behavioral expressions. Cognitive states, on the other hand, comprise complex mental processes, such as engagement [14], [15], [16], boredom [17], [18], [19], and self-confidence [20], [21], [22], which are essential for understanding an individual's mental and emotional states.

Previous research has consistently shown that self-confidence and learning are interrelated [23], [24]. Studies have shown that a significant boost in self-confidence among students can lead to substantial improvements in their learning outcomes and overall academic performance [25]. Quantified learning aims to contribute to this, bridging the uncertainty that may arise from the shift to online learning. Quantified learning would provide valuable information regarding learners' inner experiences by allowing for personalized interventions and tailored advice and encouragement when it matters. Although the relationship between eye movements and levels of self-confidence in this mode of education is not fully understood, the growing trend of telepresence indicates the need for further studies to clarify this relationship.

Building on the work of Ishimaru et al., our study extends their concept of a confidence-aware learning assistant, which uses an eye-tracker to detect self-confidence while students answer multiple-choice questions and adapt the review process based on the estimated confidence levels [21]. Our contribution extends their work by comparing two methods for estimating self-confidence based on eye-tracking data in video-based learning. Figure 1 presents an overview of the proposed method. Initially, we utilize a feature-based approach leveraging traditional classification algorithms, including Support Vector Machines (SVM) and Random Forests (RF). Subsequently, we propose a deep-learning approach based on the Long-Short-Term Memory (LSTM) network, which offers a more effective way to classify self-confidence levels and demonstrates the superiority of a deep-learning-based approach over traditional hand-crafted feature-based methods.

The main motivation behind this study is to address the growing need for personalized learning experiences that cater to the diverse needs and abilities of learners. Traditional learning systems often rely on explicit feedback mechanisms, such as self-reported confidence levels or multiple-choice question scores, which may not accurately reflect a learner's true understanding or confidence. By leveraging eye-tracking data, our study aims to develop a more nuanced and objective understanding of learner confidence, enabling the creation of more effective and adaptive learning systems.

Furthermore, this study bridges a research gap in the field of affective computing and learning analytics, where there is a need for more comprehensive and comparative studies on the use of sensor-based approaches for estimating learner confidence. By comparing the efficacy of conventional machine-learning techniques with deep-learning approaches, our study provides a complete understanding of the strengths and limitations of each method, ultimately contributing to the development of more sophisticated and effective learning systems. The research questions are as follows:

1) What are the most effective methods for incorporating eye-tracking data into video-based learning platforms to improve learning outcomes?
2) Can machine-learning-based approaches be developed to accurately assess and predict individuals' confidence levels in various educational settings?
3) How do the predictive performances of conventional feature extraction methods and deep-learning techniques compare in estimating self-confidence levels, and what are the implications for educational research and practice?

The remainder of this paper is structured as follows. Section II provides a detailed explanation of the technical background and other research that has been done on the subject of estimating confidence through sensor-based approaches. Section III describes the methodology used for gathering data. Section IV explains the methods involving eye tracking and analyzing data used to estimate learner confidence. Section V presents the results of our experiment. Section VI discusses the results in the context of our research questions and future work in this area. Finally, Section VII summarizes the main contributions of this paper and provides a conclusion.

II. RELATED WORK

In this section, we explore previous research about confidence and neurocognitive states, the use of eye-tracking in education, and the relationship between eye gaze and self-confidence.

A. CONFIDENCE AND NEUROCOGNITIVE STATES

Studies throughout academic apparatus have pondered upon the role of confidence as a factor affecting different

neurocognitive states, from standardized learning [26] to cognitive tests [27] and culinary skills [28]. Forbes-Riley and Litman found out that adding confidence to tutor systems improves learning pace and overall satisfaction [29]. At the same time, it is evident that those who learn to stand up for themselves are always successful in their endeavors. They automatically gain more confidence levels through appreciation for their performance.

According to Sun and Yeh, boosting confidence can assist students in the identification of misconceptions in their minds, where learning has become distorted, and students believe they have the correct answer, although, in reality, they are mistaken [30]. Roderer and Roebers and his colleagues' findings of age-related discrepancies hint at a possible age-related gap in self-confidence, where younger people have a greater sense of self-confidence than older people [31]. For a long, neuroscientists have discovered the linkage between physiological factors, such as EEG (electroencephalography) and self-efficacy (the person's level of confidence in performing a particular task) by exploring the relationship [30], [32].

Nevertheless, some EEG devices make partners feel bored because of their constant attachment. Instead, eye trackers provide a more convenient and native way of integrating with emergency messages on display screens. In this context, Maruichi et al. proposed a novel method to estimate a user's self-confidence based on their stroke-level handwriting behavior. This method enhances learning by enabling users to more efficiently review areas of unacquired knowledge through feedback tailored to their self-confidence [33].

Complementing the work, Bruhin et al. presented a laboratory experiment involving a team effort task where effort and ability are complementary, and synergies exist between teammates' efforts, revealing the impact of self-confidence on teamwork [34]. The study finds that overconfidence leads to increased effort, reduced free-riding, and higher team revenue when subjects' self-confidence about their ability is explicitly manipulated through easy and hard general knowledge quizzes.

B. EYE-TRACKING IN EDUCATION

Even though mobile eye-trackers have the potential to discover some learning patterns in a particular setting, the gulf is still relatively large between the very tight situations of research labs and unpredictable real-life situations. This discrepancy poses an impassable obstacle to the elaboration of correct models reflecting learning and information interaction in the real world.

To close this gap, scientists now tend to study the research in the lab and people's natural habitats. As an example highlighted by [35], it is both the case and that some have been undertaking intensive long-term studies where they captured more than 80 hours of mobile eye-tracking data. Previous research has utilized employees' commercial Electrooculography glasses to record in 27 months [36].

In line with this trend [37], provides a comprehensive review of recent eye-tracking studies within educational settings, particularly focusing on children and adolescents. It analyzes 68 empirical studies with 78 experiments emphasizing the use of eye-tracking to monitor engagement, learning interactions, and cognitive activities. The review identifies common practices in data analysis and interpretation, stressing the importance of cross-validation with other data sources. Our study aligns with this shift of attention and practical experimenting by looking at the eye-tracking feature to explore behavioral real-world learning patterns.

Eye-tracking, having great potential in this field, is also supported by the fact that it is well known that this is the area where eye movement, both proficiency in language [38] and self-esteem, are linked. Research has found that when students get stuck with a given content, their reading speed slows, and they go back to sections more often [39]. Moreover, the evidence presented by Tsai et al. shows that students' eye movements, whether from going over a question several times or looking at an explanation they understand, can be indicative of their comprehension of the concepts [40].

C. EYES AND SELF-CONFIDENCE ESTIMATION

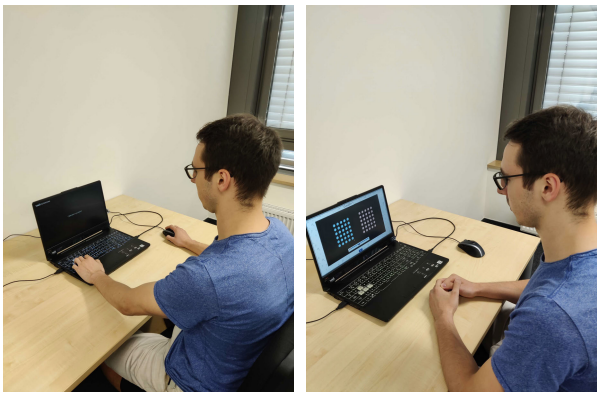
Eye behaviors and body language contribute significantly to self-confidence. Those with low self-confidence tend to spend long revising and re-evaluating every question or choice [41]. With self-assessment being only one of the many benefits of this method, eye-tracking also sheds light on enhancing the learning experience. Okoso et al. brings the idea to the fore by providing readers with information on which aspects of the text cause great difficulty in grasping the plot [42].

On the other hand, Lee et al. showed a positive correlation between eye contact with virtual tutors and an acceleration of learning [43]. Augereau et al. have carried out a good English language proficiency estimation task and achieved high accuracy with very slim error margins while using eye movements during tests [38]. Yamada et al., are among the best, representing the top exploration towards automatically identifying the self-confidence level during problem-solving using the eye movement analysis [22]. These attempts point out the potential of eye tracking to give a general idea about the learner's process and tailor feedback and learning reinforcements for students.

Through ecological validity between controlled situations and the real world, with the understanding of valuable insights from eye movements, the researcher is now opening the floodgates on comprehending the learning process. This process is a springboard for individual learning and a groundbreaking step toward the one-day development of personalized and reduced educational frameworks for all.

III. DATA COLLECTION

This section provides a detailed description of the gaze data collection process. An overview of the experimental settings and data collection workflow is presented in Figure 2. The



(a) Calibration process. (b) Participant during experiment.

FIGURE 2. Experimental setting. A participant works on the laptop with an eye-tracker mounted.

following subsections provide further information on the participant demographics and the data collection protocol.

A. PARTICIPANTS

The participant pool for our experiment consisted of 13 university students, comprising eight males and five females. The students were recruited from various academic backgrounds, including applied and theoretical science, computer science, cognitive linguistics, and mechanical engineering, thereby providing a diverse and representative sample for testing our framework.

B. PROTOCOL

The experiment was conducted using a Tobii 4C remote eye-tracker with a pro license key in a laboratory setting, isolated from potential environmental distractions. The video stimuli used for recording gaze data covered various topics, including logic, literature, computer science, and medicine, to ensure broad thematic coverage and representative eye movement and attention patterns from the participants. A series of videos were carefully selected to achieve this goal, and the data collection procedure is described in detail below.

1) An experiment conductor gave a precise description to each participant during the data collection process and the general desires that followed.
2) Every participant was required to go through the form carefully and needed to sign the consent form if they were fully aware of it and agreed with the study.
3) Participants of the research project endured a calibration procedure in which several stages were applied to achieve accurate and reliable measurements, as shown in Figure 2. This was carried out to put the eye-tracking equipment to the test and set the groundwork for all participant's individual eye-tracking and gaze patterns.
4) The participants were told to sit in front of the computer screen; now, they started watching those videos for approximately one to two minutes. The participants were prompted to follow these instructions as part of the experimental rules: To watch the videos cautiously.
5) Simultaneously to the participants viewing the video, accuracy in the quality of the pupillary data was measured in milliseconds; therefore, the eye data collection was concluded immediately upon the completion of the video.
6) Immediately after the participants watched the video, they were asked to complete a questionnaire to test their memory and understanding of the presentation. For each question, there were four choices, of which only one was the choice of the correct answer. Respondents acknowledge the correct answer by clicking on one of the choices below.
7) As soon as the participants began to write down their responses, we initiated eye data acquisition during IP. In contrast, the eye movements of the participants were recorded, and the whole session ended until they submitted their last answer to the question.
8) After participants reacted to every question with a different one, we used those to get an idea of their confidence level. The respondents filled in their answer options (Yes/No) to indicate the degree of their assurance in their answers. The confidence level reported by the participants themselves was prepared as a reference or benchmark to measure the system's performance regarding self-confidence.
9) After they indicate their confidence level by clicking the corresponding button, the following video will be played automatically.
10) The last phase was that they were asked to go ahead and continue watching videos (Steps 4-9) and to reduce them with the appropriate question after completing each video.

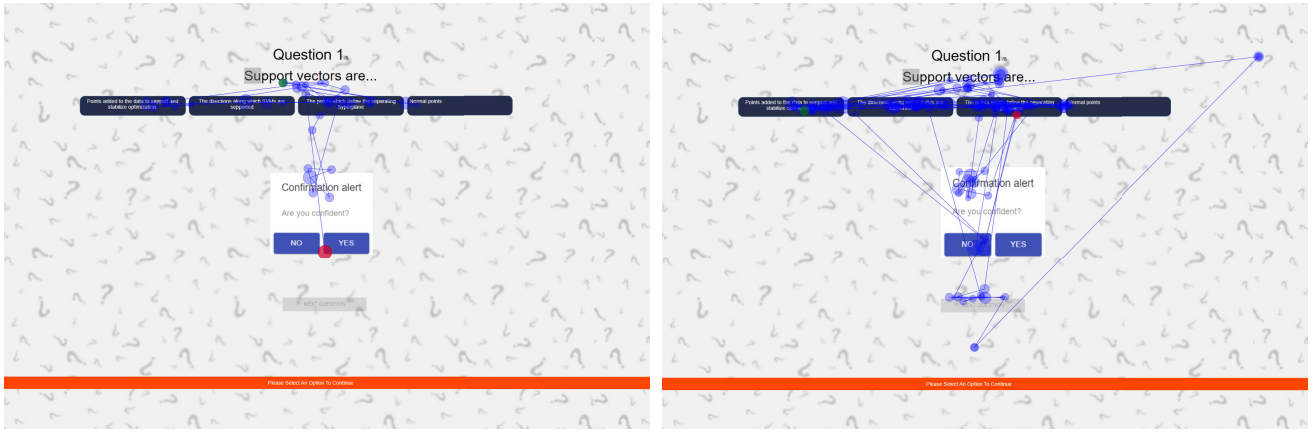
The experiment was 30 minutes long. During this trial process, the desktop used for the experiment was repeatedly fixed to keep it standing without shaking. The environment building up to the place was designed attentively so that no debris or exposure to other devices leading to any problems with recording the gaze data was of mildest concern. Moreover, the system's volume was standardized for all participants to ensure consistency, and screen brightness was set in all the experiments without any variation. This step was taken to establish an environment that was as optimal as possible and standardized so that no factor that could affect the results negatively, such as outside influences, would not hamper the data collection.

IV. METHODOLOGY

In this section, we describe the dataset preparation process, followed by the methodology for feature extraction. Subsequently, we outline the machine-learning and deep-learning models utilized. Finally, the approach to compare model accuracy.

A. DATA PRE-PROCESSING

The accuracy of eye-tracking data can be affected by various noise sources, such as blinks and head movements.



(a) Confident (b) Not confident

FIGURE 3. Example showing eye gaze when answering the questions asked with confidence and with no confidence.

To mitigate this, noise-reducing techniques are employed to eliminate distorting components and obtain more reliable eye movement indicators. This would assist in precisely examining eye movement rates, including fixations (prolonged gazes at a single location) and saccades (rapid eye movements between fixations).

In this context, fixation refers to maintaining gaze at a specific location for a brief duration (typically less than one second), while saccade is a smooth eye movement from one fixation to another. To analyze the presented model, we employed the technique developed by Buscher et al. to detect fixations and saccades [44]. Instead of exporting the absolute coordinates of fixations, we exported the differential coordinates, which capture the changes in position between consecutive fixations. The differences in eye gaze patterns while solving questions with and without confidence are illustrated in Figure 3.

B. FEATURE EXTRACTION

Our approach combines traditional feature extraction techniques with advanced deep-learning methods to provide a comprehensive evaluation of self-confidence estimation. We extracted a set of hand-crafted features from the eye-tracking data, including fixation duration, saccade length, saccade angle, and saccade speed which are detailed in Table 1 which are crucial in capturing the essential information from the data. By condensing the data into these meaningful dimensions, we aim to provide a more accurate and relevant representation of how the audiovisual materials influence participants' eye movements and understanding.

C. MODEL ARCHITECTURE

This study explores conventional hand-crafted feature-based approaches and deep neural networks to determine their effectiveness in predicting self-confidence in video-based learning environments. We employed a two-fold approach: Firstly, we leveraged the power of deep learning by utilizing

TABLE 1. The list of features.

No	Feature
$1-2$	Fixation duration {mean, std}
$3-4$	Saccade length {mean, std}
$5-6$	Saccade angle {mean, std}
$7-8$	Saccade speeds {mean, std}

Long Short-Term Memory (LSTM) neural networks, which are well-suited for modeling complex temporal relationships in data. In parallel, we adopted a more traditional approach, combining hand-crafted features with machine-learning algorithms such as Support Vector Machines (SVM) and Random Forest (RF).

1) HAND-CRAFTED FEATURE-BASED MODEL

Our hand-crafted approach employs two established machine-learning algorithms for self-confidence estimation: SVM and Random Forest. For the SVM, we used the features listed in Table 1 and selected the Radial Basis Function (RBF) Kernel as the kernel function for the SVM, as it is well-suited for handling non-linear relationships between the features. Using a grid search approach, we identified the optimal hyperparameter values as $C = 1$ and $\gamma = 0.125$, which resulted in the highest performance. We then applied the same feature set to the Random Forest algorithm with `n_estimators` $= 100$ and `criterion` $=$ "gini", which is an ensemble learning method that combines the predictions of multiple decision trees to produce a more accurate and robust prediction.

2) DEEP-LEARNING MODEL

Our deep-learning module was based on a Long Short-Term Memory (LSTM) architecture, as depicted in Figure 4.

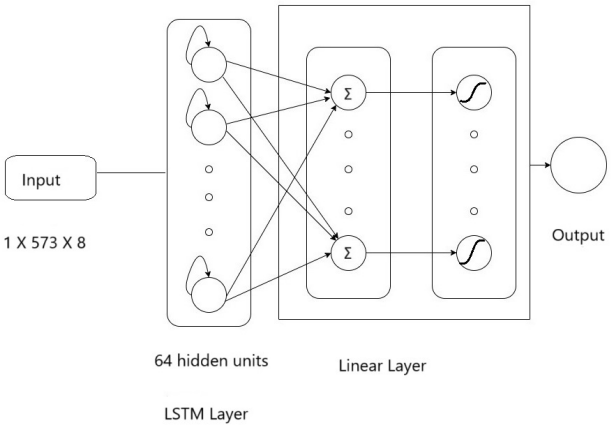


FIGURE 4. Base model architecture of LSTM.

To prepare the input data for the network, we applied padding to the data points by duplicating the data sequences, ensuring a consistent input length. As a result, the regression analysis is valid up to a maximum of 573 data points. The LSTM network consisted of a single layer with 64 hidden units, followed by a fully connected layer. The model was trained using the Adam optimizer with an initial learning rate of 0.001, and the Binary Cross-Entropy loss function was employed as the objective function.

D. EVALUATION PROTOCOL

In our study, we took advantage of two cross-validation methods, *Leave-One-Participant-Out* (LOPO) and *Leave-One-Category-Out* (LOCO), to effectively compare hand-crafted feature-based techniques and deep-learning methods. These techniques thoroughly assessed our model's performance across participants and video content.

1) LEAVE-ONE-PARTICIPANT-OUT CROSS-VALIDATION (LOPOCV)

The LOPOCV approach excluded one participant from the training set during each iteration and used their data for testing. This process was repeated until every participant had been excluded and tested once. This method allowed us to evaluate the model's ability to generalize and accurately predict self-confidence for new, unseen participants. The final accuracy was calculated as the average of all accuracies obtained from each iteration, providing a comprehensive measure of the model's overall performance across different participants.

2) LEAVE-ONE-CATEGORY-OUT CROSS-VALIDATION (LOCOCV)

The LOCOCV technique focused on the samples of watching and solving activities together. In each iteration, one category (solving and watching together) was removed from the training set and used as the test set. The model was then trained on the remaining categories. This procedure was repeated until each category had been excluded and tested once. By employing this method, we assessed the model's performance in predicting self-confidence based on different categories. Similar to LOPOCV, the final accuracy was determined by averaging the accuracies from each iteration, giving an overall performance metric for the model across different categories.

By implementing LOPOCV and LOCOCV cross-validation methods, we ensured that our technique was rigorously tested and performed well across participants and video samples. These cross-validation techniques were crucial for identifying the specific features associated with each participant and training video, thereby enhancing the reliability and generalizability of our model.

TABLE 2. Comparison of model results using LOPOCV and LOCOCV cross-validation.

Model	LOPOCV	LOCOCV
SVM	54.0%	52.0%
RF	52.0%	46.0%
LSTM	**70.3%**	**76.9%**

V. RESULT

In this section, we present the results of our comprehensive comparison between hand-crafted feature-based methods and a deep-learning-based approach. Our goal is to provide a thorough understanding of each methodology's strengths and weaknesses in the context of our research.

Table 2 presents a comparison of model results using LOPOCV and LOCOCV under two distinct approaches: a deep-learning-based method utilizing Long Short-Term Memory (LSTM) networks and hand-crafted feature-based models employing Support Vector Machines (SVM) and Random Forest (RF) algorithms.

A thorough examination of the results reveals that the deep-learning approach, leveraging the power of LSTM networks, outperforms the accuracy of hand-crafted feature-based methods. The deep-learning-based LSTM approach achieves an accuracy of 70.3% and 76.9% for LOPOCV and LOCOCV, respectively, which is significantly higher than the accuracy achieved by the SVM and RF models. The SVM model achieved an accuracy of 54.0% and 52.0% for LOPOCV and LOCOCV, respectively, while the RF model achieved an accuracy of 52.0% and 46.0% for LOPOCV and LOCOCV, respectively. These results demonstrate the efficacy of deep-learning-based approaches, particularly LSTM networks, in our research domain.

The significant performance gap between the deep-learning approach and the traditional machine-learning methods highlights the efficiency of deep-learning in handling complex data patterns and relationships. The ability of deep learning to learn and represent complex features and

TABLE 3. Results of Leave-One-Participant-Out Cross-Validation (LOPOCV) with LSTM Model.

Participant ID	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
Accuracy (%)	57.1	71.4	71.1	71.1	85.7	100.0	85.7	71.4	57.1	71.4	57.1	57.1	57.1

TABLE 4. Results of Leave-One-Category-Out Cross-Validation (LOCOCV) with LSTM Model.

Video ID	V1	V2	V3	V4	V5	V6	V7
Accuracy (%)	61.5	69.2	76.9	92.3	84.6	69.2	84.4

patterns in the data allows it to achieve higher accuracy and better generalization performance than traditional machine-learning methods. The deep-learning approach demonstrates a significant percentage increase in accuracy compared to the traditional machine-learning methods in the LOCOCV approach. Specifically, the deep-learning approach achieved an accuracy of 76.9%, which represents a 47.5% increase over the SVM model's accuracy of 52.0% and a 67.4% increase over the RF model's accuracy of 46.0%. These results highlight the substantial improvement in accuracy achieved by the deep-learning approach compared to the machine-learning methods.

Table 3 presents each participant's Leave-one-participant-out cross-validation (LOPOCV) results. The model's accuracy varies significantly across participants, ranging from 57.1% to 100.0%. *Participant 6* achieved the highest accuracy of 100.0%, indicating that the model was able to predict their behavior perfectly. In contrast, *Participants 1, 9, 12, and 13* achieved lower accuracy values of 57.1%, suggesting that the model struggled to predict their behavior. The model's performance is inconsistent across all participants, highlighting the need for further research to improve the model's performance and generalizability. The results suggest that the model can accurately predict the behavior of some participants but not others and that individual differences in behavior might explain this.

Leave-One-Category-Out Cross-Validation (LOCOCV) experiment results using an LSTM model are presented in Table 4. The accuracy of the model varies across categories (video samples), ranging from 61.5% to 92.3%. The model achieved the highest accuracy of 92.3% for *Video 4*, indicating that it was able to predict the behavior for this sample accurately. In contrast, the model achieved lower accuracy values for *Video 1 and 6*, with accuracy values of 61.5% and 69.2%, respectively.

Figure 5 presents the confusion matrices for the LOPOCV of the SVM, RF, and LSTM models. Different scenarios illustrate the binary classification model's performance in predicting confidence levels during question-solving and video-watching activities. Figure 5a, Figure 5b, and Figure 5c depict the results when inferences are made in a user-dependent manner for the SVM, RF, and LSTM models, respectively.

Figure 6 presents the confusion matrices for LOCOCV of the SVM, RF, and LSTM models. Figure 6a, Figure 6b, and Figure 6c show the results when inferences are made in a video-dependent manner for the same models. These confusion matrices provide a comprehensive visual representation of the model's performance, allowing for a detailed comparison of their predictive capabilities across different validation approaches and inference contexts.

VI. DISCUSSION

This section provides a discussion and interpretation of the results presented in Section V. The discussion examines the implications of the findings and addresses the research questions posed in Section I.

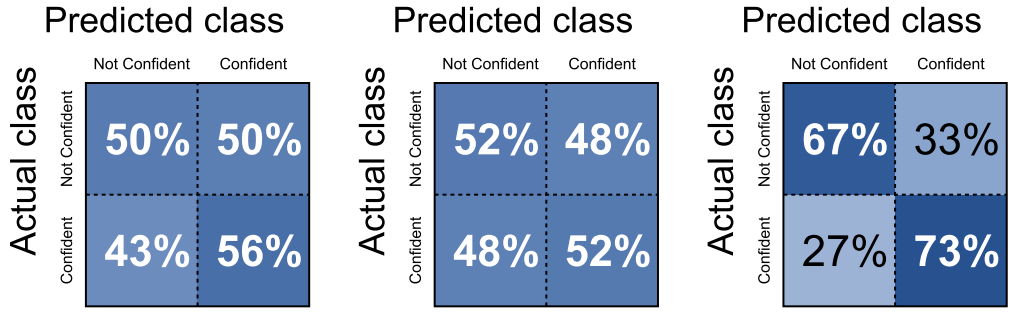
A. INCORPORATING EYE-TRACKING DATA INTO VIDEO-BASED LEARNING PLATFORMS

We have introduced a data collection procedure incorporating eye-tracking data from video-based learning scenarios. This comprehensive dataset includes valuable information such as confidence levels, raw gaze data, and correctness of responses. It provides a robust foundation for analyzing and estimating self-confidence in educational contexts.

The data collection approach is unique in its generalizability and applicability across educational contexts. To ensure data generalizability and alignment potential, all videos are carefully selected to represent a wide range of variety of topics at multiple levels of difficulty. We further ensured diversity as participants were recruited from a variety of ages, backgrounds, and learning preferences using an extensive participant recruitment process. The wide range of background information aids in the generalized as well as reliability, rendering our dataset a useful asset for forthcoming research works on educational psychology and technology-supported learning.

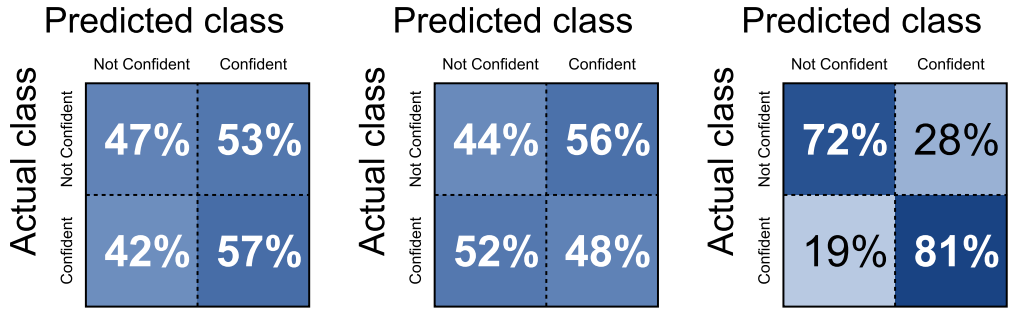
B. CONVENTIONAL MACHINE-LEARNING TECHNIQUES FOR SELF-CONFIDENCE ESTIMATION

Our approach involved an extensive examination of individuals' confidence levels using conventional machine-learning methods. Specifically, we utilized Support Vector Machines (SVM) and Random Forests (RF) to estimate confidence levels from the collected data. These methods provided



(a) LOPOCV with SVM model (b) LOPOCV with RF model (c) LOPOCV with LSTM model

FIGURE 5. Confusion matrices of SVM, RF, and LSTM models for LOPOCV.



(a) LOCOCV with SVM model (b) LOCOCV with RF model (c) LOCOCV with LSTM model

FIGURE 6. Confusion matrices of SVM, RF, and LSTM models for LOCOCV.

a baseline for evaluating the effectiveness of traditional machine-learning techniques in predicting self-confidence based on eye-tracking data.

There are several reasons that we have selected to implement SVM and RF. Firstly, SVM is used across industries for being efficient in high-dimensional spaces and its ability to construct hyperplanes that optimally separate classes in a feature-rich dataset. It progressively supports the use case of handling eye-tracking data. Furthermore, in cases with fewer points of data compared with the number of dimensions, SVM has a better generalization scale than other methods (for us, this was discussed in our dataset). RF is an ensemble learning method that works by constructing a multitude of decision trees at training time and reporting the mode of classes (classification) or mean prediction (regression) per-tree outputs. This method delivers better accuracy by combining multiple models, which eliminates the overfitting problem. In addition, since RF has no restrictions on the number of features it can manage, this is consistent with our complex and varied eye-tracking data nature.

SVM and RF require carefully engineered features to perform well. So, considerable manual effort and domain expertise are required to figure out which parts of the raw eye-tracking data are likely key to predicting self-confidence. Important as it is, feature engineering can be very time-consuming and serve to limit the portability of these types of models between data sources. In addition, traditional models such as SVM and RF may not be able to model the temporal dynamics in eye-tracking data as well. Eye movements are inherently spatiotemporal, and the order and timing of gazes can provide crucial information about cognitive and emotional states. Traditional machine-learning models are not inherently suited to cope with such temporal dependencies, leading them to underperform when deployed.

C. COMPARISON OF CONVENTIONAL HAND-CRAFTED FEATURE-BASED APPROACH TO DEEP-LEARNING-BASED APPROACH

We conducted a comparative analysis of manually designed feature extraction methods versus deep-learning techniques for predicting self-confidence levels. We employed LSTM neural networks for deep learning, which demonstrated superior performance compared to traditional machine-learning methods. The LSTM model outperformed SVM and RF, achieving higher accuracy in both user-dependent and video-dependent scenarios. This finding underscores the potential of deep-learning models to uncover latent patterns in eye-tracking data, leading to more accurate and reliable predictions of self-confidence levels.

D. LIMITATIONS AND FUTURE WORK

While this study on self-confidence estimation using deep learning provides valuable insights, several limitations and challenges must be acknowledged. These include the

potential variability in model accuracy and the complexity of interpreting eye-tracking data.

Firstly, the reliance on eye-tracking devices, such as the Tobii 4C, underlines a dependency on specific hardware for data collection, which may limit the applicability of the models in environments where such equipment is unavailable. Additionally, variations in the accuracy and calibration of eye trackers can affect the reliability of the collected data.

The assessment of self-confidence is limited to responses related to video content, which may not be representative of all aspects of self-confidence, as it is influenced by a wide range of factors and contexts. Self-confidence is a multifaceted trait influenced by numerous internal and external factors, and focusing on a specific scenario may oversimplify the broader concept and its determinants.

The study has been conducted with a limited or homogeneous sample population, which may affect the generalizability of the findings. A diverse sample is crucial to ensure that the models can accurately estimate self-confidence across different demographics and cultural backgrounds.

For future work, it is important to consider the variety of online lecture formats, such as those involving teacher-led discussions or document-based materials. Different types of learning materials can impact learners' self-assessments. Further exploration could provide insight into the influences of factors such as age, prior knowledge, learning styles, and personality on the estimation models.

Integrating eye-tracking with additional sensors remains an important task for future research. Physiological measures, such as EEG and biofeedback, can provide deeper insights into cognitive and emotional states. Additionally, incorporating vocal intonations, facial expressions, and eye movements could lead to a more precise understanding of participants' confidence levels.

Future work could focus on developing more complex models that help explain the factors behind confidence estimates, enhancing our understanding of learner behavior and model transparency. These models could be applied in personalized learning platforms that offer tailored feedback, targeted interventions, and customized learning paths that adapt to individual needs. It will be challenging to evaluate their performance in authentic educational settings with many students, but it is essential for widespread adoption.

VII. CONCLUSION

We conducted an experiment where computer models based on deep learning could rate self-confidence using eye-tracking data. Participants watched videos and answered questions with their gaze movements captured. In a comparison of different models, the LSTM achieved an average prediction accuracy of 73.6% for the degree of self-confidence, outperforming Support Vector Machines and Random Forests, which had accuracies around 53.0% and 49.0%. This finding also hints that LSTM models can provide feedback and support in domains such as education, where strengthening confidence and skills is vital. Additional research would provide testing of the models' generalizability, handling multimodal data integration, and developing interventions based on these models' insights. This research shows us that we can utilize deep learning to understand and motivate self-confidence.

REFERENCES

[1] M. Ciotti, M. Ciccozzi, A. Terrinoni, W.-C. Jiang, C.-B. Wang, and S. Bernardini, "The COVID-19 pandemic," *Crit. Rev. Clin. Lab. Sci.*, vol. 57, no. 6, pp. 365–388, Jul. 2020.

[2] S. Leo, N. M. Alsharari, J. G. Abbas, and M. Alshurideh, "From offline to online learning: A qualitative study of challenges and opportunities as a response to the COVID-19 pandemic in the UAE higher education context," in *The Effect of Coronavirus Disease (COVID-19) on Business Intelligence*, Jan. 2021, pp. 203–217.

[3] S. J. Daniel, "Education and the COVID-19 pandemic," *Prospects*, vol. 49, nos. 1–2, pp. 91–96, Apr. 2020.

[4] R. Higashimura, K. Watanabe, A. Vargo, M. Iwata, A. Dengel, and K. Kise, "Estimating unknown English words from user smartphone reading behaviors," *IEEE Access*, vol. 12, pp. 140223–140234, 2024.

[5] T. Mizumoto, Y. Otoda, C. Nakajima, M. Kohana, M. Uenishi, K. Yasumoto, and Y. Arakawa, "Design and implementation of sensor-embedded chair for continuous sitting posture recognition," *IEICE Trans. Inf. Syst.*, vol. E103.D, no. 5, pp. 1067–1077, 2020.

[6] L.-H. Lee and P. Hui, "Interaction methods for smart glasses: A survey," *IEEE Access*, vol. 6, pp. 28712–28732, 2018.

[7] H. Suzawa, K. Watanabe, M. Iwamura, K. Kise, A. Dengel, and S. Ishimaru, "Supporting smooth interruption in a video conference by dynamically changing background music depending on the amount of utterance," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.* New York, NY, USA: Association for Computing Machinery, Sep. 2022, pp. 299–302, doi: 10.1145/3544793.3560384.

[8] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Trans. Audio, Speech Language Process.*, vol. 15, no. 8, pp. 2190–2201, Nov. 2007.

[9] K. Watanabe, Y. Soneda, Y. Matsuda, Y. Nakamura, Y. Arakawa, A. Dengel, and S. Ishimaru, "DisCaaS: Micro behavior analysis on discussion by camera as a sensor," *Sensors*, vol. 21, no. 17, p. 5719, Aug. 2021.

[10] C. Chen, Y. Arakawa, K. Watanabe, and S. Ishimaru, "Quantitative evaluation system for online meetings based on multimodal microbehavior analysis," *Sensors Mater.*, vol. 34, no. 8, p. 3017, 2022.

[11] T. Hayashida, Y. Nakamura, H. Choi, and Y. Arakawa, "Privacy-aware quantitative measurement of psychological state in meetings based on non-verbal cues," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops Other Affiliated Events (PerCom Workshops)*, Mar. 2024, pp. 433–436.

[12] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Toward practical smile detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2106–2111, Nov. 2009.

[13] H. Liu, Y. Gao, and P. Wu, "Smile detection in unconstrained scenarios using self-similarity of gradients features," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1455–1459.

[14] 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.

[15] K. Watanabe, A. Dengel, and S. Ishimaru, "Metacognition-EnGauge: Real-time augmentation of self-and-group engagement levels understanding by gauge interface in online meetings," in *Proc. Augmented Humans Int. Conf.* New York, NY, USA: Association for Computing Machinery, Apr. 2024, pp. 301–303, doi: 10.1145/3652920.3653054.

[16] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," 2016, *arXiv:1609.01885*.

[17] G. N. Yannakakis, H. P. Martínez, and A. Jhala, "Towards affective camera control in games," *User Model. User-Adapted Interact.*, vol. 20, no. 4, pp. 313–340, Oct. 2010.

[18] D. Giakoumis, D. Tzovaras, K. Moustakas, and G. Hassapis, "Automatic recognition of boredom in video games using novel biosignal moment-based features," *IEEE Trans. Affect. Comput.*, vol. 2, no. 3, pp. 119–133, Jul. 2011.

[19] S. M. Feraru and M. D. Zbancioc, "Emotion recognition for disgust and boredom states," in *Proc. Int. Symp. Signals, Circuits Syst. (ISSCS)*, Jul. 2017, pp. 1–4.

[20] 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.

[21] S. Ishimaru, T. Maruichi, A. Dengel, and K. Kise, "Confidence-aware learning assistant," 2021, *arXiv:2102.07312*.

[22] K. Yamada, K. Kise, and O. Augereau, "Estimation of confidence based on eye gaze: An application to multiple-choice questions," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. ACM Int. Symp. Wearable Comput.*, Sep. 2017, pp. 217–220.

[23] I. Sadler, "The role of self-confidence in learning to teach in higher education," *Innov. Educ. Teaching Int.*, vol. 50, no. 2, pp. 157–166, May 2013, doi: 10.1080/14703297.2012.760777.

[24] J. C. Ortiz-Ordoñez, F. Stoller, and B. Remmele, "Promoting self-confidence, motivation and sustainable learning skills in basic education," *Proc. Social Behav. Sci.*, vol. 171, pp. 982–986, Jan. 2015.

[25] E. A. Linnenbrink and P. R. Pintrich, "The role of self-efficacy beliefs instudent engagement and learning intheclassroom," *Reading Writing Quart.*, vol. 19, no. 2, pp. 119–137, Apr. 2003.

[26] R. Jersakova, R. J. Allen, J. Booth, C. Souchay, and A. R. O'Connor, "Understanding metacognitive confidence: Insights from judgment-of-learning justifications," *J. Memory Lang.*, vol. 97, pp. 187–207, Dec. 2017.

[27] S. Kleitman and J. Gibson, "Metacognitive beliefs, self-confidence and primary learning environment of sixth grade students," *Learn. Individual Differences*, vol. 21, no. 6, pp. 728–735, Dec. 2011.

[28] J. A. Pooler, R. E. Morgan, K. Wong, M. K. Wilkin, and J. L. Blitstein, "Cooking matters for adults improves food resource management skills and self-confidence among low-income participants," *J. Nutrition Educ. Behav.*, vol. 49, no. 7, pp. 545–553, Jul. 2017.

[29] K. Forbes-Riley and D. Litman, "Adapting to student uncertainty improves tutoring dialogues," in *Proc. AIED*, Jul. 2009, pp. 33–40.

[30] J. C.-Y. Sun and K. P.-C. Yeh, "The effects of attention monitoring with EEG biofeedback on university students' attention and self-efficacy: The case of anti-phishing instructional materials," *Comput. Educ.*, vol. 106, pp. 73–82, Mar. 2017.

[31] T. Roderer and C. M. Roebers, "Can you see me thinking (about my answers)? Using eye-tracking to illuminate developmental differences in monitoring and control skills and their relation to performance," *Metacognition Learn.*, vol. 9, no. 1, pp. 1–23, Apr. 2014.

[32] L.-H. Hsia, I. Huang, and G.-J. Hwang, "Effects of different online peer-feedback approaches on students' performance skills, motivation and self-efficacy in a dance course," *Comput. Educ.*, vol. 96, pp. 55–71, May 2016.

[33] T. Maruichi, S. Ishimaru, and K. Kise, "Self-confidence estimation on vocabulary tests with stroke-level handwriting logs," in *Proc. Int. Conf. Document Anal. Recognit. Workshops (ICDARW)*, vol. 3, Sep. 2019, pp. 18–22.

[34] A. Bruhin, F. Petros, and L. Santos-Pinto, "The role of self-confidence in teamwork: Experimental evidence," *Experim. Econ.*, vol. 27, no. 3, pp. 687–712, Jul. 2024.

[35] J. Steil and A. Bulling, "Discovery of everyday human activities from long-term visual behaviour using topic models," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2015, pp. 75–85.

[36] S. Ishimaru, K. Hoshika, K. Kunze, K. Kise, and A. Dengel, "Towards reading trackers in the wild: Detecting reading activities by EOG glasses and deep neural networks," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, Sep. 2017, pp. 704–711.

[37] F. Ke, R. Liu, Z. Sokolikj, I. Dahlstrom-Hakki, and M. Israel, "Using eye-tracking in education: Review of empirical research and technology," *Educ. Technol. Res. Develop.*, vol. 72, no. 3, pp. 1383–1418, Jun. 2024.

[38] O. Augereau, H. Fujiyoshi, and K. Kise, "Towards an automated estimation of English skill via TOEIC score based on reading analysis," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1285–1290.

[39] K. Rayner, "Eye movements in reading and information processing: 20 years of research.," *Psychol. Bull.*, vol. 124, no. 3, pp. 372–422, 1998.

[40] M.-J. Tsai, H.-T. Hou, M.-L. Lai, W.-Y. Liu, and F.-Y. Yang, "Visual attention for solving multiple-choice science problem: An eye-tracking analysis," *Comput. Educ.*, vol. 58, no. 1, pp. 375–385, Jan. 2012.

[41] K. Kojima, K. Muramatsu, and T. Matsui, "Experimental study toward estimation of a learner mental state from processes of solving multiple choice problems based on eye movements," in *Proc. 20th Int. Conf. Comput. Educ. (ICCE)*, Jan. 2012.

[42] A. Okoso, T. Toyama, K. Kunze, J. Folz, M. Liwicki, and K. Kise, "Towards extraction of subjective reading incomprehension: Analysis of eye gaze features," in *Proc. 33rd Annu. ACM Conf. Extended Abstr. Hum. Factors Comput. Syst.*, Apr. 2015, pp. 1325–1330.

[43] H. Lee, Y. Kanakogi, and K. Hiraki, "Building a responsive teacher: How temporal contingency of gaze interaction influences word learning with virtual tutors," *Roy. Soc. Open Sci.*, vol. 2, no. 1, Jan. 2015, Art. no. 140361.

[44] G. Buscher, A. Dengel, and L. van Elst, "Eye movements as implicit relevance feedback," in *Proc. CHI Extended Abstr. Hum. Factors Comput. Syst.*, Apr. 2008, pp. 2991–2996.

ANKUR BHATT was born in New Delhi, India, in 1998. He received the B.Tech. degree in computer science from Manav Rachna University, Faridabad, India, in 2020. He is currently pursuing the master's degree in computer science specializing in artificial intelligence with the University of Kaiserslautern-Landau (appointed by the University of Kaiserslautern, in 2021, and its campus was merged with the University of Koblenz and Landau, in 2023). He is also a Student Research Assistant with the Department of Smart Data and Knowledge Services, German Research Center for Artificial Intelligence GmbH (DFKI), Kaiserslautern. His current research interests include gaze estimation, eyewear computing, machine translation, and natural language processing.

KO WATANABE was born in Hiroshima, Japan, in 1994. He received the B.E. degree in mechanical engineering from Tokyo University of Agricultural and Technology, Japan, in 2017, and the M.E. degree from Nara Institute of Science and Technology, Japan, in 2019. He has been a Ph.D. Researcher with German Research Center for Artificial Intelligence (DFKI), since 2021, and is a working member of the Immersive Quantified Learning Laboratory (IQL Laboratory) and Smart Data and Knowledge Services (SDS) Department, DFKI Kaiserslautern. He was a Software Engineer with DeNA, Tokyo, Japan. His current research interests include technologies that augment human intellect and explainable AI.

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 has been a Ph.D. Researcher with German Research Center for Artificial Intelligence (DFKI), since 2019, and is a working member of the Immersive Quantified Learning Laboratory (IQL Laboratory) and Smart Data and Knowledge Services (SDS) Department, DFKI Kaiserslautern. In addition, he has been a Teaching Assistant with RPTU Kaiserslautern-Landau, in 2022. He was a Research Assistant with DFKI, from 2017 to 2019, and has published papers at IEEE Access, Ubicomp, and IUI conferences. His research interests include deep-learning-based affective and mental 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).

ANDREAS DENGEL received the Diploma degree in CS from TUK and the Ph.D. degree from the University of Stuttgart. He is currently the Scientific Director of DFKI GmbH, Kaiserslautern. In 1993, he became a Professor in computer science with TUK, where he holds the Chair of Knowledge-Based Systems. Since 2009, he has been appointed as a Professor (Kyakuin) with the Department of Computer Science and Information Systems, Osaka Prefecture University. He was with IBM, Siemens, and Xerox Parc. He is a member of several international advisory boards, has chaired major international conferences, and founded several successful start-up companies. He is a co-editor of international computer science journals and has written or edited 12 books. He is the author of more than 300 peer-reviewed scientific publications and supervised more than 170 master's and Ph.D. theses. He is a fellow of IAPR and received many prominent international awards. His main scientific emphasis is in the areas of pattern recognition, document understanding, information retrieval, multimedia mining, semantic technologies, and social media.

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 Project Professor with the Graduate School of Informatics, Osaka Metropolitan University, Japan, since 2023. In addition, he has been a Researcher with the Keio Media Design Research Institute, since 2014. He was a Junior Professor with the University of Kaiserslautern-Landau, Germany, from 2021 to 2023, and has been a Senior Researcher with German Research Center for Artificial Intelligence (DFKI), Germany, since 2019. His research interests include human–computer interaction, machine learning, and cognitive psychology with the aim of amplifying human intelligence. His awards and honors include the Best Presentation Award at the Asian CHI Symposium 2020, Poster Track Honorable Mention at UbiComp/ISWC 2018, and 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.