Deep Learning and Sensor-Driven Learning Analytics and Augmentation

Thesis approved by the Department of Computer Science
RPTU University Kaiserslautern-Landau
for the award of the Doctoral Degree
Doctor of Engineering (Dr.-Ing.)
to

Jayasankar Santhosh

Date of Defence: 18.11.2025
Dean: Prof. Dr. Christoph Garth (RPTU Kaiserslautern-Landau)
Reviewers: Prof. Dr. Prof. h.c. Andreas Dengel (RPTU Kaiserslautern-Landau),
Prof. Dr. Koichi Kise (Osaka Metropolitan University)

DE 386

Abstract

Enhancing e-learning outcomes requires understanding students' learning behaviors, particularly their affective and cognitive states. By recognizing and responding to these states, personalized interventions can be implemented to improve learning efficacy. This research aims to develop adaptive, learner-centric systems that integrate real-time mental state predictions with customized support, promoting both academic performance and student well-being in digital learning environments.

To address the challenges in detecting learners' mental states, this research focuses on developing innovative deep-learning solutions that integrate multimodal sensor data, enabling more accurate predictions and personalized interventions to enhance e-learning outcomes. We have developed and implemented innovative deep learning based models incorporating various validation methods and techniques to improve generalized prediction and personalized prediction models with significant improvements in overall accuracies. Notably, fine-tuning and user-specific calibration resulted in a substantial performance increase, with three-level stress detection using an LSTM model achieving a 56% improvement to reach 91% accuracy (F1=0.911) and a person-specific CNN-LSTM model showing a 50% improvement in detecting interest levels per participant. User-dependent approaches further enhanced accuracy, yielding gains of 13% for engagement, 19% for arousal, and 15% for valence.

These refined models provide the foundation for personalized interventions designed to enhance the learning experience. These interventions include an adapted cognitive control training for managing distractions during online learning where the findings indicated that participants who received training exhibited significantly improved comprehension (*p-value = 0.0059*) and reduced distraction levels (*p-value = 0.021*) compared to the untrained group. Additionally, an application-based feedback system provides visual cues based on the learner's current engagement and emotional state, enabling students to self-regulate and adapt their learning strategies in real time.

Building on these interventions, a gaze-based adaptive learning system dynamically adjusts content presentation based on real-time engagement analysis. This system utilizes AI-generated summaries via ChatGPT to provide concise and relevant information when low engagement is detected, helping to re-engage students and maintain focus. User studies demonstrated significant improvements in comprehension ($p = 0.0018$), engagement ($p = 0.0021$), and overall learning outcomes, further highlighting the effectiveness of these personalized approaches.

Acknowledgements

I am deeply grateful to Prof. Andreas Dengel, for his invaluable guidance and support throughout my research. I sincerely thank him for providing me with the opportunity to work at the German Research Center for Artificial Intelligence (DFKI). His insightful comments and suggestions have greatly contributed to the success of my work.

I would like to express my heartfelt gratitude to Prof. Shoya Ishimaru, for his constant encouragement and thoughtful guidance throughout my PhD jorney. His expertise and constructive feedback have been instrumental in shaping my research and achieving my academic goals.

I would like to sincerely thank Dr. Nicolas Großmann for his thoughtful coordination and support at the Immersive Quantified Learning Lab. I am also deeply grateful to Brigitte Selzer, Dr. Thomas Kieninger, Dr. Ko Watanabe, Dr. David Dzsotjan and for their valuable advice. Their insightful comments and guidance were always helpful whenever I encountered challenges.

I would like to sincerely express my gratitude to my colleagues at DFKI : Soumy Jacob, Javier Carrasco Melo, Steffan Steinert, Haruki Suzawa, Kanta Yamaoka, Dai Shimzu, Yuichiro Iwashita, Riku Higashimura and Tokio Uchida. Their collaboration, support, and encouragement have made my research journey both productive and enjoyable.

Most importantly, thanks to my students: Pramod Vadiraja, Prerna Garg, Rashmi Alur Ramachandra, Akshay Pai, Hanan Moulay, Shrishti Jagtap, Shokouh Zadeh Shoushtari, Ankur Bhatt, Gitesh Gund, Taufik Tamboli and Suramya Chandra. Their dedication, hard work, and invaluable support played a crucial role in the successful completion of this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid expansion of e-learning platforms has transformed educational practices, reshaping how teaching and learning are approached [50, 202]. However, traditional e-learning systems often ignore learners' affective and cognitive states, leading to decreased motivation, increased stress, and reduced learning outcomes. Recent advancements in deep learning and sensor technologies, such as eye-tracking and physiological sensing, present new opportunities to analyze and detect learning behaviors more effectively [84]. For example, Ishimaru et al. [86] proposed that matching instructional material with the cognitive states of students might improve their understanding and learning outcomes. Despite these advancements, the lack of robust cognitive and affective state detection systems and personalized interventions continues to limit the potential of e-learning to deliver high-quality, adaptive educational experiences.

This research aims to address these limitations by developing innovative deep-learning solutions that leverage sensor data, including eye-tracking, physiological signals, and their combinations, to enhance learning outcomes. By integrating these data sources, the proposed approach provides a comprehensive understanding of learners' mental and emotional states, enabling accurate predictions and personalized interventions [160]. These technologies allow for the identification of critical factors such as interest, stress, engagement, distractions, and emotional states within the learning environment. By accurately detecting and responding to learners' affective and cognitive states, the system can deliver personalized support that enhances engagement, improves comprehension, and reduces stress. This not only boosts academic performance but also fosters a more supportive and effective learning environment. Through this research, the main aim is to unlock the full potential of e-learning platforms, equipping them to meet the diverse needs of learners in an evolving educational landscape.

1.1 Basic Concepts

The primary emphasis in the field of learning analytics has been on examining the behaviors that learners exhibit while utilizing digital tools [54]. Typically, these digital tools consist of intelligent tutoring systems [133], learning management systems [184], online courses, educational games, or various other platforms that involve a computer as a key element in the learning journey. Conversely, significantly fewer studies in learning analytics have been carried out in alternative educational settings, such as collaborative in-person environments where computers are either absent or play a secondary role. This inclination toward digital learning environments is effectively illustrated by the need for a learning analytics system to document a record of the educational experience. The integration of deep learning and sensor-based learning analytics has further transformed e-learning by providing personalized and adaptive educational experiences by enabling the real-time analysis of learners' cognitive and affective states, allowing for tailored interventions that enhance the learning experience [31, 81].

This research proposed the enhancement of digital learning through three interconnected phases: detection, intervention, and adaptation. This approach involves detection of affective and cognitive states of learners using innovative deep learning models, the utilization of the model predictions for designing personalized interventions, and using real-time sensor data and effective model predictions for dynamic adaptation of learning materials.

1.1.1 Detection of Learner States

The ability to accurately detect and interpret learner states is fundamental to creating effective personalized learning environments, reflecting the insight of educational psychologist Benjamin Bloom who noted, "What any person in the world can learn, almost all persons can learn if provided with appropriate prior and current conditions of learning" [19]. Deep learning has emerged as a significant factor in enhancing cognitive and affective computing, reshaping the way machines perceive and react to human emotions and cognitive conditions [160]. The complex dynamics of human feelings and thoughts, such as facial expressions, vocal variations, and physiological signals, can be efficiently analyzed by deep learning models [160]. Particularly, Transformer networks, Autoencoders, Convolutional Neural Networks (CNNs) [206], and Recurrent Neural Networks (RNNs) [70], along with their advanced iterations, have shown considerable promise in interpreting the complex temporal and spatial relationships found in emotional and cognitive data [160]. The integration of various sensors offers a detailed perspective on how learners engage with educational materials. D'Mello et al. [42] provided an extensive review of multimodal affect detection, focusing on methods for merging physiological and behavioral data, with the insights gained potentially applicable for developing sophisticated reading engagement and emotion recognition systems [160]. The research conducted by D'Mello and Graesser highlighted the significant impact of changing affective states on complex learning [47]. Schmidt et al. made a noteworthy advancement by introducing a new, publicly accessible dataset that comprises multimodal sensor data on stress and emotion, gathered from wearable devices across different activities [162]. The authors emphasize the dataset's potential to foster the development of new algorithms and methodologies for ongoing, real-time stress and emotion recognition.

1.1.2 Designing Personalized Interventions

The idea of personalized interventions in education is based on the recognition that each learner is distinct, reflecting the thoughts of psychologist Carl Rogers, who said, "The only person who is educated is the one who has learned how to learn and change" [151]. This viewpoint emphasizes the necessity of customizing educational experiences to meet individual requirements. In e-learning, personalized interventions use insights gained from detecting emotional and cognitive states to develop adaptable learning settings. As noted by Pardo et al. [140], these interventions strive to deliver "the right content to the right learner at the right time," thereby enhancing the learning experience. Recent advancements in learning analytics and artificial intelligence have significantly enhanced our ability to design and implement such interventions effectively. Karumbaiah et al. [96] found that adaptive interventions tailored to students' emotional states in an intelligent tutoring system led to a 15% improvement in problem-solving efficiency and a significant increase in positive affect. A study by Taub et al. [181] revealed that personalized prompts based on learners' self-regulated learning behaviors in a game-based learning environment resulted in a 30% increase in learning gains compared to a control group without such interventions. These developments mark a significant shift towards more responsive and learner-centric educational approaches.

1.1.3 Dynamic Adaptation of Learning Materials

To further optimize the learning environment building on personalized interventions, the dynamic adaptation of learning materials is the next step [64, 178, 199]. This approach continuously adjusts content in real-time based on the learner's requirements and mental states. By using real-time sensor data and deep learning models, adaptive learning systems can make instant decisions about content presentation, difficulty levels, and learning pathways. For example, Zhang et al. primarily utilized deep learning models to tailor content presentation based on individual learning patterns, resulting in enhanced engagement and improved learning outcomes [203]. Arroyo et al. [9] investigated how learning materials can be dynamically adapted by utilizing real-time sensor information and deep learning techniques to develop tailored and responsive educational experiences that enhance student cognition, engagement, and emotional response. They attribute this enhancement to particular elements and interventions that are fundamentally affective, cognitive, and metacognitive. These advancements in dynamic adaptation can create learning experiences that are both personalized and responsive to the learner's changing state, making education more effective and engaging.

1.2 Research Questions

This thesis addresses three key research questions. Understanding learners' affective and cognitive states for optimizing e-learning is important to developing systems supporting personalized learning. For that, the thesis came up with three research questions to answer.

1.) How can sensor data and deep learning be utilized to effectively detect learners' affective and cognitive states? Detecting learners' mental states is significant in understanding their needs and preferences. One question is how effectively deep learning based models can estimate learners' states using multimodal sensor data.

2.) How to design and deliver personalized interventions that effectively address learners' mental states and improve their learning experience? Designing tailored interventions responding to the detected learner states is significant to understand the success of the learning process. The question here addresses how to design and deliver personalized interventions that could impact the learning process of a learner.

3.) How to adapt learning materials to learners' real-time mental states? As a last question, the thesis aims to answer if an adaptive learning system can dynamically adjust the learning materials to meet learners' changing needs and states, improving the overall learning experience.

1.3 Contributions

In summary, contributions of this thesis include:

- Development and evaluation of novel deep learning models for affective and cognitive state detection in e-learning, including advanced calibration techniques for improving generalized predictions.
- Design and implementation of innovative interventions for mitigating digital reading distractions.
- Development of a visual feedback system based on sensor data and affective/cognitive states to enhance user experience.
- Design and implementation of a gaze-based adaptive learning system that integrates AI-generated adaptive content delivery based on real-time user states.

The research contributions make a significant impact on understanding, modeling, and influencing individual learning behavior. The study investigates the potential of various deep learning approaches and sensor fusion techniques to enhance the accuracy of predicting learners' affective and cognitive states. The work thoroughly examines and refines the computation of cognitive and affective states, achieving robust results for both generalized and personalized validation scenarios. The personalized intervention techniques and adaptive learning system developed have the potential to transform the e-learning ecosystem, creating a smooth balance between the learner and the environment. By understanding and catering to each individual's unique needs, the research empowers learners to reach their full potential, making learning a more engaging, effective, and enjoyable journey. By creating an environment that feels personalized and supportive, the approach helps reduce stress and increase motivation, ultimately leading to improved learning outcomes and a pleasant learning experience.

1.4 Outline of Thesis Chapters

Chapter 2 provides an overview of affective and cognitive computing, deep learning for state detection, personalized intervention strategies, and adaptive learning systems. Chapter 3 presents various approaches for detecting affective and cognitive states in learners. Chapter 4 discusses personalized intervention strategies to improve cognitive performance and presents a real-time gaze-based adaptive learning system. Finally, Chapter 5 provides a conclusion to the thesis by summarizing its key aspects and outlining future directions.

Chapter 2

Background and Related Work

This chapter provides a review of prior work on the key concepts, technologies, and methodologies that form the basis of this research. It begins with an overview of affective and cognitive computing 2.1, followed by an exploration of related work on deep learning for state detection 2.2, emphasizing its advantages, scalability, and personalization capabilities. The approaches for designing personalized interventions in e-learning are discussed in Section 3. Lastly, Section 4 focuses on adaptive learning systems that respond dynamically to learners' real-time needs.

2.1 Affective and Cognitive Computing

Stress detection has been studied using subjective and objective methods. Subjective approaches include self-reported questionnaires like the Perceived Stress Scale (PSS) [33], Daily Stress Inventory (DSI) [22], and Brief Symptom Inventory (BSI) [37]. Objective methods rely on physiological signals from wearable sensors. EEG and ECG are widely used to measure stress through brain activity and heart rate variability [2, 40]. Ahn et al. used EEG and ECG features to detect stress during cognitive tasks, achieving 87.5% accuracy with SVM models [2]. Other signals, such as electrodermal activity (EDA), blood volume pulse (BVP), and skin temperature, are also effective [145, 196]. Gjoreski et al. developed a multimodal stress classification framework using wearable data, achieving up to 92% accuracy in real-life scenarios [61]. Cho et al. combined PPG and thermal imaging to create a smartphone-based stress monitoring system using neural networks [32].

Engagement detection measures attention and interest during learning. Conati et al. used eye-tracking to study how students respond to adaptive hints in educational games, showing that the timing and design of hints significantly impact attention and learning [34]. Jacob et al. used wristbands to detect reader interest, while Wang et al. conducted a review on multi-sensor eye-tracking systems that are used to measure attention [89, 191]. Ishimaru et al. discovered links between pupil size, nasal temperature, and cognitive states such as engagement with educational materials [85]. Other studies have used multimodal approaches to monitor engagement and detect distractions [6, 27, 113, 134].

Emotion detection is critical for understanding learners' affective states. D'Mello and Graesser examined how affective states influence complex learning, demonstrating their effect on the learning process [48]. Schmidt et al. presented a multimodal dataset aimed at detecting stress and emotions via wearable devices, aiding the creation of real-time emotion recognition systems [162]. Calvo and D'Mello conducted a review of methods for affect detection, emphasizing issues such as the intricacy of emotions and the variability of input data [26]. Arroyo et al. implemented emotion sensors in educational contexts to tailor content according to students' emotional reactions, although practical obstacles persist [8] [160].

Readability assessment quantifies the difficulty of text for readers. Early methods, such as those by Kincaid et al., relied on shallow features like sentence length and syllable count [55, 154]. Schwarm et al. combined traditional metrics with statistical language models and parsers [165]. Most research focuses on native speakers, but Heilman et al. showed that grammatical features can predict reading difficulty for both first and second language learners [73]. Feng et al. developed readability-prediction algorithms for automatic text simplification, advancing the field [53].

Despite significant advancements, challenges remain in affective and cognitive computing. Stress detection often struggles with variability in physiological signals across individuals and contexts. Engagement and emotion detection face difficulties in real-time accuracy and the integration of multimodal data. Readability assessment lacks robust models for non-native speakers and diverse linguistic contexts. This research aims to address these gaps by developing adaptive, multimodal frameworks that improve accuracy, personalization, and scalability in real-world learning environments.

2.1.1 Eye-Tracking and Physiological Sensing while Learning

Eye-tracking technology and physiological sensors have been widely utilized to gather data and infer readers' cognitive and affective states in learning environments. Eye tracking, in particular, has been employed to monitor eye movements and gaze patterns during reading, providing insights into attention, comprehension, and interest [36, 97, 167]. Recent studies have also explored the use of eye tracking in immersive virtual reality (VR) and augmented reality (AR) learning environments, highlighting its potential to decode attention and cognitive states in dynamic settings [93, 112]. Physiological sensing techniques, including measures like heart rate, skin conductance, and facial expressions, have also been used to capture readers' emotional and cognitive responses during learning [80]. These signals provide valuable insights into arousal, valence, and emotional engagement, enabling a deeper understanding of learners' states and their interaction with educational content. Asteriadis et al. estimated user behavior in an e-learning environment by analyzing eye gaze and head pose detection to determine whether a user was frustrated, distracted, tired, or fully engaged [10]. Similarly, Ishimaru et al. focused on detecting attention and cognitive states in learning scenarios, particularly in the context of human-document interaction using eye-tracking and physiological sensing [83].

Physiological data analysis has significant potential for revealing subtle emotional variations, including shifts in heart rate, skin conductance, or facial expressions, which are closely related to students' engagement and cognitive activities [143]. Alongside physiological signals, eye-tracking data is crucial in identifying engagement and emotions during reading tasks [176]. Combining gaze data with physiological signals provides specific benefits in recognizing emotion and engagement while reading [122]. Gaze data offers essential insights into a reader's visual attention and concentration on particular aspects of the text, whereas physiological signals, such as heart rate and skin conductance, provide objective assessments of emotional arousal and valence [115, 177]. Brishtel et al. 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 [24]. Brishtel et al. highlighted the strong potential of the EDA sensor for detecting mind wandering, and the integration of electrodermal activity with eye-tracking features significantly improved the precision of mind wandering classification [24] [160].

2.1.2 Multimodal Sensor Integration and Fusion Techniques

The integration of multiple sensors for detecting cognitive and affective states provides a rich and detailed understanding of how learners interact with text [4]. Research in multimodal affective analysis [172, 192] often focuses on various strategies for combining data from different modalities, typically categorized into early fusion (at the feature level), late fusion (at the decision level), model-level fusion, and hybrid approaches. Koelstra et al. significantly advanced this area by introducing a specialized dataset designed for analyzing human affective states [101]. Their work highlighted the value of incorporating physiological signals into the analysis, suggesting that understanding emotions through such measures could greatly improve the quality and effectiveness of reading and learning interventions. D'Mello et al. offered a detailed review of multimodal affect detection, including methods for combining physiological and behavioral data, with insights that could be applied to develop advanced systems for recognizing reading engagement and emotion [43]. Chen et al. developed a model for detecting learner engagement by fusing video, text, and physiological signals, and their findings indicated that this model could accurately assess learning engagement [29]. In their study, Tzirakis et al. utilized a CNN for extracting features from audio and a deep residual network for visual data [186]. Hao et al. presented a combined visual-audio emotion detection system that employed multitask and blending learning techniques with diverse features [66]. This system used SVM classifiers and CNNs for both handcrafted and deep learning-based visual-audio features, resulting in four sub-models, and achieved average accuracies of 81.36% (speaker-independent) and 78.42% (speaker-dependent) on the eNTERFACE dataset [117].

Advances in wearable technologies have increased interest in automated affective analysis using multiple physiological modalities [164]. However, due to the complex nature of emotions and the considerable variation in individual physiological responses [205], achieving reliable prediction results with EEG-based or ECG-based emotion detection remains challenging. In studies focusing on physiological modality fusion for affective analysis, various types of physiological signals (e.g., ECG, EOG, BVP, GSR, and EMG) are often combined with EEG and ECG to interpret emotional states [109]. The visual modality (such as facial expressions, voice, gestures, and posture) can also be integrated with multimodal physiological signals for a more comprehensive visual-physiological affective analysis [195]. Verma and Tiwary [188] adopted a thorough approach to emotion estimation by leveraging a range of physiological data, including ECG, GSR, ST, BVP, RESP, EMG, and EOG, which were selected from the DEAP dataset. Their methodology involved using the Discrete Wavelet Transform (DWT) [72] to extract multi-resolution features from this data. These features were then crucial for estimating emotions across three dimensions: valence, arousal, and dominance [160].

2.2 Deep Learning for State Detection

Deep learning emerged as a strong instrument in cognitive and affective computing, enhancing machines' capacity to comprehend and react to human emotions and cognitive states. It offers effective means to analyze complex human behaviors, including facial expressions, vocal inflections, and physiological signals. Notably, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), along with their more advanced forms, have demonstrated their utility in capturing the intricate temporal and spatial relationships present in emotional and cognitive data [71, 207].

Martinez et al. introduced an end-to-end deep learning framework specifically for affect recognition using physiological signals. Their methodology integrated denoising autoencoders, CNNs, preference learning, and automated feature selection. This framework was applied to the Maze-ball dataset, which comprises BVP and EDA data collected during gameplay, supplemented by corresponding questionnaires [98]. In a related effort, Qiu et al. developed Correlated Attention Networks for emotion detection, which utilized bidirectional Gated Recurrent Units (GRUs) in conjunction with a Canonical Correlation Layer, a Signal Fusion Layer, an Attention Layer, and a Classification Layer. Their model was evaluated on three prominent datasets: SEED, SEED IV, and DEAP, demonstrating its effectiveness in emotion recognition tasks [147] [160].

Badshah et al. developed a deep CNN with three convolutional and fully connected layers to extract features from spectrogram images, enabling the prediction of seven emotions [12]. Zhang et al. enhanced this approach using a pre-trained AlexNet DCNN with a discriminant temporal pyramid matching (DTPM) strategy to capture global utterance-level features [204]. Dziezyc et al. compared ten deep learning models for affect recognition across four datasets, finding CNNs more effective than LSTMs for physiological sensor data [45]. Santamaria-Granados demonstrated the potential of DCNNs for emotion detection using the AMIGOS dataset, highlighting their effectiveness in affective computing [123, 155]. Watanabe et al. used MobileNetV2 to assess student engagement during online meetings, achieving an F1-score of 89.5% with leave-one-participant-out cross-validation [193].

While deep learning has made significant advancements in state detection, many models struggle with generalizability across diverse datasets and real-world scenarios due to variations in physiological signals, environmental factors, and individual differences. The reliance on large labeled datasets for training deep learning models poses challenges, as collecting and annotating such data is time-consuming and resource-intensive. Real-time processing and scalability also remain significant challenges, particularly in applications requiring low-latency responses, such as engagement detection in online learning environments. This research aims to address these challenges by developing adaptive deep learning frameworks that integrate multimodal data, improve generalization across diverse contexts, and reduce dependency on labeled datasets. The focus on lightweight architectures could enable real-time applicability in practical learning scenarios.

2.2.1 Scalability, Performance, and Personalization

Scalability, performance, and personalization are key considerations in the application of deep learning to state detection in cognitive-affective computing. This is very significant because such systems are being increasingly deployed in real-world applications such as education, health, and human-computer interaction; hence, these systems have to be efficient, adaptive, and relevant to users. Scalability is one of the main challenges because most deep learning models proposed for state detection require high computational resources both for training and inference. Lightweight architectures, such as MobileNet and EfficientNet, have been proposed in the literature for reducing computational complexity while maintaining high accuracy [79, 180]. Edge computing and cloud-based solutions have also been explored in order to provide scalable deployment of the state detection system in resource-constrained environments [171].

Performance is critical for real-time state detection in applications such as engagement monitoring or emotion recognition since these applications need low-latency and high-accuracy predictions. Model pruning, quantization, and knowledge distillation are among several techniques used to optimize deep learning models for faster inferences with minimum compromise on accuracy [65, 77, 87]. In addition, there has also been much development of hardware accelerators, like GPUs and TPUs, which have greatly enhanced the performance of deep learning models for real-world applications [94].Personalization provides a gain in accuracy and relevance of state detection systems, since there is much individual variation in emotional and cognitive states. Personalization may be performed by using transfer learning, where pre-trained models are fine-tuned on user-specific data, or adaptive learning frameworks that update in the light of user feedback [106, 139]. It also spawned the creation of federated learning, a promising approach wherein models learn from decentralized user data in a private manner [121].

This research addresses the development of lightweight adaptive deep learning frameworks by fusing multi-modal data in improving computational efficiency with a grant toward personalized state detection.

2.3 Personalized Interventions in E-Learning

User applications, feedback systems, and visualization tools have become integral to personalized interventions in e-learning, leveraging advancements in affective computing and human-centered design. Personalized digital learning systems enable tutors to tailor their instruction to individual learners' requirements and learner characteristics [82, 126]. Martin et al. highlighted that adaptive feedback is among the most extensively used and commonly implemented strategies for adaptation in digital learning [116]. Researchers suggest that adaptive and personalized feedback can enhance student performance [129], address the developmental and motivational needs of secondary students [102], and foster self-regulated learning [137].

For instance, Shaikh et al. introduced a tool to help social science experts analyze emotional patterns influenced by cultural stressors, providing insights into shifts in emotional responses [168]. Watanabe et al. developed EnGauge, an application that visualizes user engagement during online meetings through a gauge interface, offering real-time feedback on attention levels [193]. Similarly, Grad-CAM-based heatmaps [166] have been employed to highlight critical features in input facial images, enhancing the interpretability of CNN-based models. Selvaraju et al. further refined this approach with Guided Grad-CAM, combining high-resolution visualizations with class-discriminative localization to improve transparency in deep learning models [166]. Abdul et al. introduced COGAM, a tool designed to regulate cognitive load in interpretable AI systems through visual segmentation, advancing beyond traditional metrics of simplicity and accuracy [1]. Narciss et al. proposed a model for designing informative tutoring feedback in multimedia learning, emphasizing its function (cognitive, meta-cognitive, motivational), content (evaluative, hints, explanations), and presentation (timing, adaptivity) [128]. Bimba et al. analyzed empirical and non-empirical studies and classified feedback adaptiveness as means, targets, goals, and strategy [15].

These tools and techniques are paving the way for more personalized, adaptive, and empathetic e-learning environments, enabling interventions that cater to individual learners' cognitive and emotional needs. This research aims to build on these advancements by developing more dynamic and context based systems that integrate real-time multimodal data to deliver tailored and effective interventions.

2.4 Adaptive Learning Systems and Environments

Adaptive learning has undergone a huge evolution, transitioning from basic personalized content delivery to sophisticated, AI-driven systems that dynamically adjust to individual learner requirements [5, 108]. The contemporary adaptive learning platforms utilize advanced algorithms and artificial intelligence to modify instructional content, delivery methods, and pace in real-time, based on a learner's performance and engagement [44, 52, 62, 95]. Hussain et al. showcased cutting-edge performance with their innovative multi-layer topic modeling approach, which integrates the Felder-Silverman learning style model with fuzzy logic and sentiment analysis to accurately identify and adapt to diverse learning styles [81]. The incorporation of emotional and physiological data has emerged as a pivotal development in adaptive learning systems. Wei et al. made use of eye-tracking technology to detect learners' interest and emotional states [194], while Tyng et al. provided neuroimaging evidence for the crucial role of emotions in learning and memory [185]. Sargazi et al. developed an AI-driven decision framework that identifies and implements micro-break activities based on learners' emotional states, thereby demonstrating enhanced learning performance through emotional regulation [161].

Recent advancements in adaptive learning technologies emphasize the integration of deep learning and neural networks. Omar et al. demonstrate how neural networks can improve learning management systems by generating customized learning paths and facilitating real-time adjustments to content based on learner interactions [135]. A thorough review by Essa et al. underscores the increasing inclination towards utilizing machine learning for intelligent, adaptive e-learning landscapes, while also pointing out the necessity for additional comparative analyses of deep learning techniques [51].

Digital learning is undergoing a transformation due to Large Language Models (LLMs), which can produce text that resembles human writing and offer tailored educational experiences [208]. These models exhibit advanced skills in creating personalized learning resources, addressing student inquiries, and providing immediate explanations [92, 125, 142]. Recent developments in the use of LLM's for educational purposes have yielded promising outcomes across different fields. In programming education, Gabbay et al. showcased that GPT-4 outperforms traditional automated test-based feedback tools in generating feedback on coding assignments [57]. Azaiz et al. observed that GPT-4 Turbo delivers more organized and consistent feedback in programming courses, although they highlighted some issues with consistency [11] [156].

There are still challenges that remain in creating adaptive learning systems that could integrate cognitive, emotional, and behavioral data, providing personalized learning experiences. This research addresses these gaps by developing frameworks that utilizes deep learning, real-time data analysis, and large language models (LLMs) to dynamically adapt content, feedback, and interventions. This work aims to enhance learner engagement, improve knowledge retention, and foster self-regulated learning in diverse educational contexts.

Chapter 3
Affective and Cognitive State Detection

This chapter presents various approaches for detecting affective and cognitive states, which are essential for developing intelligent systems that can understand and respond to human emotions and needs [144]. The goal of this research is to create a comprehensive framework using various deep learning based architectures and conventional feature-extraction based approach for detecting and analyzing human states, including stress, interest, and readability.

The chapter is organized into different sections, each focusing on a specific aspect of affective and cognitive state detection. The first section presents stress detection methods using public datasets, which provide a foundation for understanding the physiological and behavioral signals associated with stress [58]. The second section explores stress detection in controlled environments, where the conditions are carefully manipulated to elicit specific stress responses [148].

The next section delves into multimodal interest detection, which involves analyzing multimodal sensor signals, such as gaze and physiological data, to infer a person's level of interest while reading carefully selected articles [90, 157]. This is followed by a section on deep learning-based emotion and engagement detection, which leverages the power of neural networks to recognize patterns in human behavior and emotional responses [138, 160].

Finally, the chapter concludes with a section on readability detection, which aims to determine the ease or difficulty of reading a particular text based on various linguistic and cognitive factors [158]. This research has significant implications for education, communication, and human-computer interaction, where the ability to detect and respond to human states can greatly enhance the effectiveness and user experience of various systems and applications.

The following sections of this chapter are based on collaborative research with students, where each contributed to specific aspects of the work. Section 3.1 presents research developed in collaboration with Prerna Garg, who assisted with data analysis and modeling using a publicly available dataset. Section 3.2 is based on work conducted jointly with Rashmi Alur Ramachandra, who contributed to data collection, supported the analysis, and application development. Similarly, Section 3.4 includes research carried out with Akshay Palimar Pai, who was involved in data collection, analysis, and application development. My own contributions span the conceptualization of the studies, formulation of research questions, technical data analysis, overall supervision of the collaborative efforts, and preparation and publication of the resulting scientific articles..

3.1 Stress Detection Using Public Datasets

Stress refers to the psychological and biological processes during emotionally and cognitively demanding situations. It is identified as one of the top ten social determinants of health disparities [58]. Organizations such as the World Health Organisation, the American Psychological Association, and the Occupational Safety and Health Administration are raising awareness about the negative impact of stress on health and its associated costs to society [58, 114].

There are several parameters that imply stress levels in the body in a scientific context, which are electrodermal activity (EDA), heart rate variability (HRV), blood volume pulse (BVP), electroencephalogram (EEG), and respiratory activity [59]. Although beneficial in the short term, prolonged stress exposure is a known contributor to several serious health conditions, such as hypertension and coronary artery disease [39]. Beyond physical health, sustained stress can also lead to mental health challenges like depression, anxiety disorders, and burnout. Therefore, it is crucial for individuals to be aware of stressful situations so they can take appropriate measures to manage them. This necessitates the development of a biofeedback system that can accurately and promptly identify stress, providing individuals with the information needed for effective intervention [58].

The main goal of this work is to provide an overview of stress detection methods on various sensor-based physiological data and investigate the publicly available multimodal physiological datasets for stress detection. We have used two datasets namely WESAD (Wearable Stress and Affect Detection) [163] and SWELL-KW [103], for evaluating our stress prediction method by employing different machine learning algorithms. The physiological data including ECG, EDA, EMG, RESP, TEMP, and also motion-based ACC data sampled at 700 Hz from a chest-worn device was used for training the models for WESAD dataset [58]. Six different machine learning models (five traditional machine learning models with manual feature extraction and one deep neural network based approach) were used for stress prediction and their performances were compared. For SWELL dataset, the HRV data [131] from SWELL-KW was used for predicting different stress conditions by using multiple machine learning classifiers.

3.1.1 Stress Detection Using WESAD Dataset

The contents of this section are based on our published work titled "*Stress Detection in Students using Wearable Sensors and Deep Learning*" [58]. The first dataset we used for analysis was WESAD, a multimodal physiological dataset which includes data from a chest-worn device (RespiBan Professional) and a wrist-worn device (Empatica E4). The various modalities measured from both chest and wrist-worn devices include electrocardiogram (ECG), electrodermal activity (EDA), electromyogram (EMG), respiration (RESP) and temperature (TEMP), blood volume pulse (BVP) and accelerometer (ACC) data [58].

The study protocol was implemented on 15 graduate students with the aim of eliciting three different affective states — neutral, stress, and amusement.

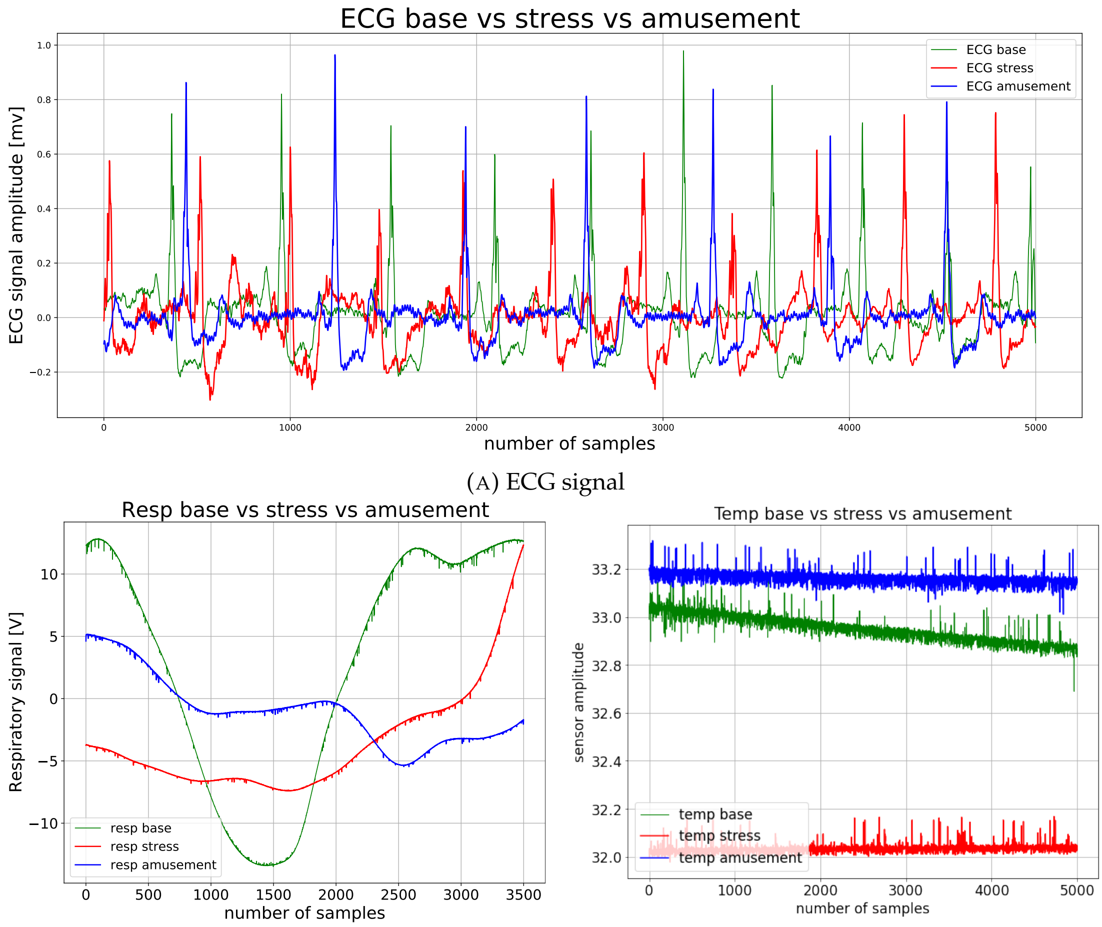
- **Baseline**: The aim of this condition was to induce a neutral affective state where the participants were reading materials by sitting or standing at a table and the recording lasted for 20 minutes.
- **Stress**: The participants were put through Trier Social Stress Test [100] where they underwent public speaking and mental arithmetic tasks. Such kind of tasks could inflict a high cognitive load on the participants and that was the intention. The participants were asked to give a five minute speech in front of a panel and after that, they were asked to count down from 2023 to zero by a factor of 17.
- **Amusement**: For this condition, the participants watched eleven funny video clips, where there was a short neutral condition of five seconds after each clip. The length of this condition was 392 seconds.

After both stress and amusement conditions, it was followed by a guided meditation period which aimed at de-exciting the participants and bringing them back to a neutral state. The participants were asked to fill out five self-report questionnaires throughout the experiment to validate the study protocol. The various questionnaires used were PANAS (Positive and Negative Affect Schedule) [183], STAI (State-Trait Anxiety Inventory) [175], SAM (Self-Assessment Manikins) [21]. An SSSQ (Short Stress State Questionnaire) [74] was also added following the stress condition.

Methodology

The WESAD dataset was utilized for stress detection using both chest-worn and wrist-worn devices, each providing unique insights into physiological signals. Both the chest-worn and wrist-worn Empatica E4 device data were analyzed using a two-fold approach, incorporating both traditional manual feature-extraction based approach and deep learning techniques, allowing for a comprehensive comparison of the performance of these approaches in detecting stress from physiological signals.

Manual Feature Extraction-Based Approach: The physiological data acquired using the chest-worn device was sampled at a frequency of 700 Hz, while the wrist-worn Empatica E4 device was sampled at different sampling frequencies: BVP (64 Hz), EDA (4 Hz), TEMP (4 Hz), and ACC (32 Hz). The raw physiological signals collected from chest-worn device during the experiments are illustrated in Figures 1a, 1b, and 1c, showing characteristic patterns of ECG, respiratory activity, and temperature variations respectively for a participant. The preprocessed and normalized sensor signals from both devices were segmented using a sliding window algorithm with a window size of 10 seconds and 50% overlap.



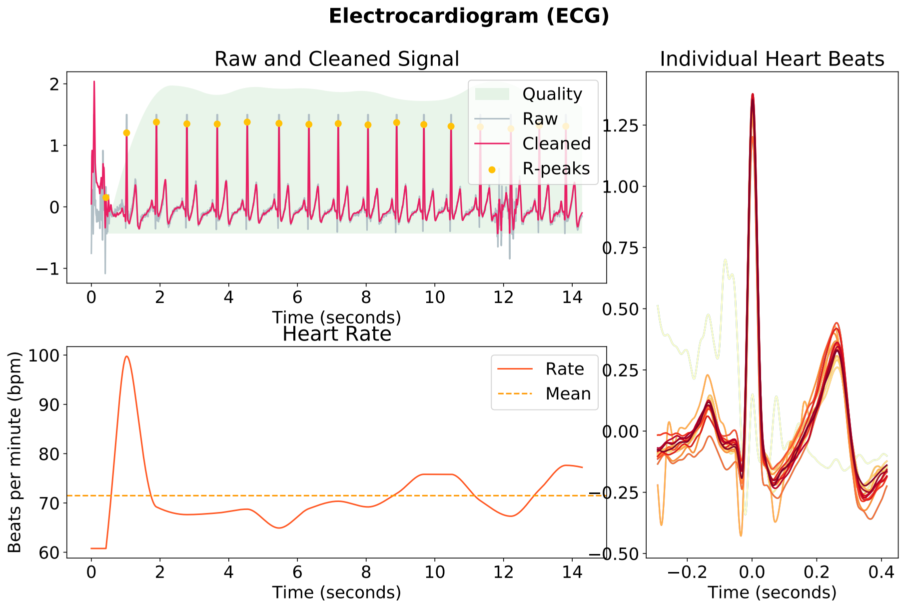
(A) ECG signal

(B) RESP signal

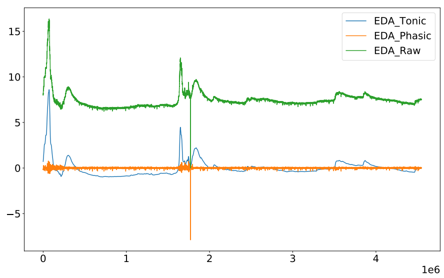
(C) Temp signal

FIGURE 1: Raw physiological signals recorded from the chest-worn device showing variations in ECG, respiratory, and temperature patterns for a participant

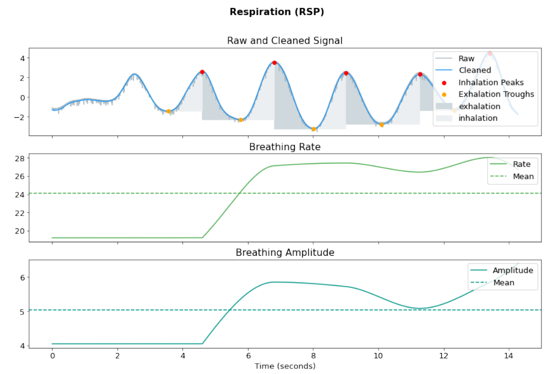
The statistical features, such as mean, standard deviation, minimum, and maximum, were extracted from the raw ACC signal for each axis for both the chest-worn and wrist-worn devices. From the raw ECG signal of the chest-worn device, the Heart Rate (HR) and R-Peaks were extracted after signal cleaning and noise removal, as illustrated in Figure 2a, and the corresponding statistical features were computed. The raw EDA signal was decomposed into its phasic (SCR) and tonic (SCL) components as shown in Figure 2b, and the statistical features were computed from these components. The respiration rate and amplitude were extracted from the cleaned RESP signal of the chest-worn device as shown in Figure 2c, and the statistical features were calculated. For the EDA, BVP, and TEMP signals of the wrist-worn device, the statistical features such as mean, standard deviation, minimum, and maximum were computed. Additionally, the peak frequencies of the BVP signal and the slope of the TEMP signal were computed for the wrist-worn device. Similarly, for the TEMP signal of the chest-worn device, all statistical features were calculated. Table 3.1 lists all the features extracted from chest data and Table 3.2 lists the features extracted from wrist data.



(A) Raw ECG and cleaned signal with individual heartbeats



(B) Raw EDA along with phasic and tonic components



(C) Raw RESP and cleaned signal with respiration rate and amplitude

FIGURE 2: The features extracted from ECG, EDA and RESP signals

Five different machine learning algorithms - Light Gradient Boosting (LightGBM), Random Forest (RF), Extra Trees (ET), Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA) - were used, and their relative performance was compared. Two types of classifications were performed: three-class (neutral vs. stress vs. amusement) and binary classification (stress vs. non- stress). The hyperparameters of the different algorithms were tuned to achieve the best possible results. For the Random Forest classifier, the minimum number of samples for splitting a node was set to five, and the number of estimators was set to 100. For SVM, the radial basis function kernel was used, and the class weight was set to 'balanced'. The ground truths in the dataset used were neutral, stress, and amused encoded as labels zero, one, and two, respectively [58].

TABLE 3.1: List of manually extracted features for chest sensor data

Signal	Features	Description
ACC	$\mu_{ACC,i}$, $\sigma_{ACC,i}$ $min_{ACC,i}$, $max_{ACC,i}$ $i \in \{x, y, z\}$	Mean, SD Min, Max Separate for x, y, z axes
ECG	HR, R-peaks μ_{HR}, σ_{HR}, min_{HR}, max_{HR} $\mu_{R-peaks}$, $\sigma_{R-peaks}$, $min_{R-peaks}$, $max_{R-peaks}$	Heart Rate, R-peak timings Mean, SD, Min, Max of HR Mean, SD, Min, Max of R-peaks
EDA	SCR, SCL μ_{SCR}, σ_{SCR}, min_{SCR}, max_{SCR} μ_{SCL}, σ_{SCL}, min_{SCL}, max_{SCL}	Phasic, Tonic components Mean, SD, Min, Max of SCR Mean, SD, Min, Max of SCL
RESP	Respiration Rate, Amplitude μ_{RR}, σ_{RR}, min_{RR}, max_{RR} μ_{Amp}, σ_{Amp}, min_{Amp}, max_{Amp}	Rate, Amplitude values Mean, SD, Min, Max of Resp Rate Mean, SD, Min, Max of Resp Amp
TEMP	μ_{TEMP}, σ_{TEMP} min_{TEMP}, max_{TEMP}	Mean, SD Min, Max

Deep Learning Based Approach: In addition to the machine learning classifiers, a deep learning model based on Fully Convolutional Network (FCN) [46] was implemented for stress detection. When compared to the conventional machine learning classifiers, the main advantage of using an FCN model is that it could learn high-level features from raw sensor signals, and manual feature extraction is not required, which requires domain expertise.

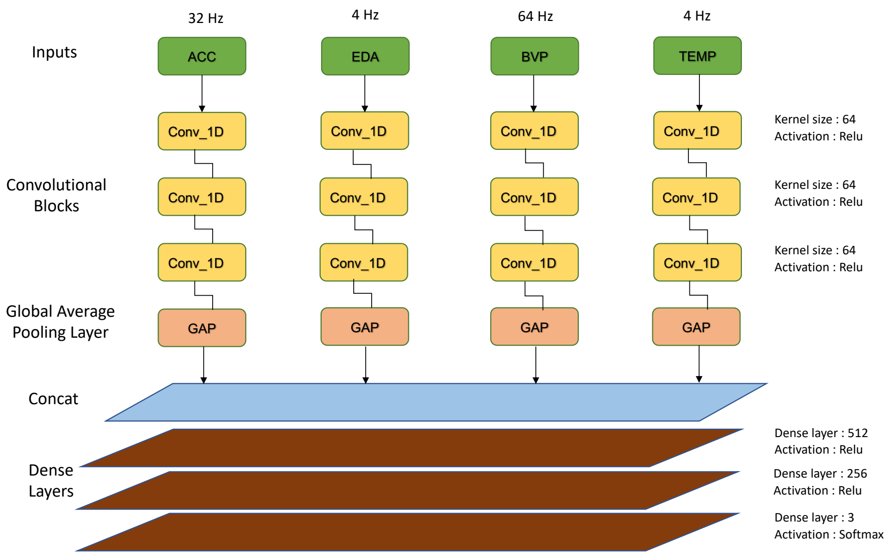


FIGURE 3: Multichannel fully convolutional network architecture

The FCN consists of three convolutional blocks for each signal, and the branches are concatenated and fed to a fully connected dense layer. The model architecture

TABLE 3.2: List of manually extracted features for wrist sensor data

Signal	Features	Description
ACC	$\mu_{ACC,i}$, $\sigma_{ACC,i}$ $min_{ACC,i}$, $max_{ACC,i}$ $i \in \{x, y, z, 3D\}$	Mean, SD Min, Max Separate for x, y, z axes
BVP	μ_{BVP}, σ_{BVP} min_{BVP}, max_{BVP} $peak_{\omega_{BVP}}$	Mean, SD Min, Max Peak frequency of BVP
EDA	μ_{EDA}, σ_{EDA} min_{EDA}, max_{EDA}	Mean, SD Min, Max
TEMP	μ_{TEMP}, σ_{TEMP} min_{TEMP}, max_{TEMP} m_{TEMP}	Mean, SD Min, Max Slope of TEMP

is shown in Figure 3. For classification, we used the raw sensor signals from both the chest-worn and wrist-worn devices. The signals were preprocessed, normalized, and segmented using a sliding window of length 10 seconds with 50 percent overlap. The hyperparameters of the convolutional blocks were tweaked to achieve the best classification performance.

Evaluation Metric: During the study protocol, various conditions were carried out at different lengths, and this led to high imbalance in the dataset. To overcome this limitation, the majority classes were under-sampled, and the minority classes were oversampled using the SMOTEENN [28, 107] technique, which was used to balance the dataset. Due to the dataset's imbalance, accuracy was not considered the most suitable evaluation metric; instead, the F1-score was utilized. To evaluate the models' ability to generalize, a leave-one-participant-out cross-validation method was implemented. The overall accuracy was determined by averaging the testing accuracies from each fold, where one participant's data was held out for testing, and the rest were used for training. This approach was specifically chosen to verify the model's performance on new, unobserved participants [58].

Results

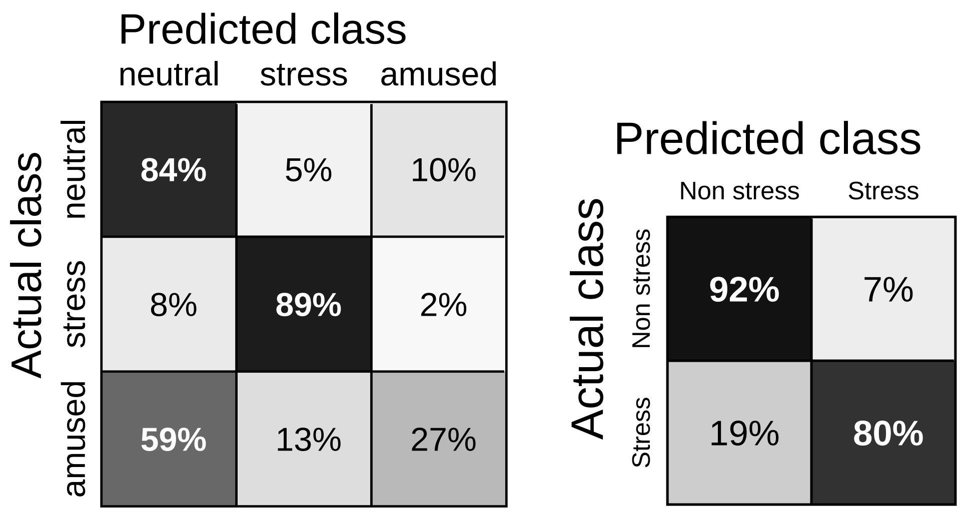
The performance of various stress detection models was evaluated using the WESAD dataset, considering signals from chest-worn, wrist-worn, and combined sensors. Both binary (stress vs. non-stress) and three-class (neutral vs. stress vs. amusement) classification tasks were performed. The models compared included traditional machine learning approaches based on manual feature extraction (SVM, LDA, Extra Trees, Random Forest) and a deep learning approach using a Fully Convolutional Network (FCN).

TABLE 3.3: Performance comparison of different models across the signals for stress detection

Model	Signals	Binary		Three-class	
		Accuracy	F1-score	Accuracy	F1-score
SVM	Chest	74.70	78.47	68.24	66.06
	Wrist	69.50	69.10	61.42	60.15
	Combined	72.88	71.64	67.22	67.04
LDA	Chest	80.20	80.04	68.52	66.98
	Wrist	74.52	72.43	60.35	57.76
	Combined	79.88	78.10	65.20	64.52
Extra Trees	Chest	83.23	82.89	65.72	63.72
	Wrist	82.55	81.12	67.64	65.80
	Combined	84.72	83.08	68.26	67.90
Random Forest	Chest	86.19	85.95	73.87	71.25
	Wrist	84.18	84.32	68.27	68.05
	Combined	87.03	86.76	72.88	72.52
FCN	Chest	91.54	89.39	**80.23**	**78.47**
	Wrist	86.42	85.29	73.91	71.66
	Combined	**91.97**	**91.35**	78.61	76.75

Among the manual feature extraction methods, Random Forest consistently achieved the highest performance across all signal sources and classification tasks. For binary classification using combined signals, Random Forest obtained 87.03% accuracy and 86.79% F1-score, with comparable performance using wrist signals (84.18% accuracy, 84.32% F1-score) and combined signals (86.19% accuracy, 85.95% F1-score). A similar trend was observed for three-class classification, with Random Forest outperforming other manual feature-extraction methods.

The deep learning-based FCN demonstrated superior performance across all signal sources, significantly outperforming manual feature extraction-based approaches. For binary classification, FCN achieved remarkable results with chest signals (91.54% accuracy, 89.39% F1-score), wrist signals (86.42% accuracy, 85.29% F1-score), and showed the best performance with combined signals (91.97% accuracy, 91.35% F1-score). The same pattern was observed in three-class classification task, where FCN achieved 80.23% accuracy and 78.47% F1-score with chest signals and maintained the performance with wrist signals (73.91% accuracy, 71.66% F1-score) and combined signals (78.61% accuracy, 76.75% F1-score). The confusion matrices using FCN for three-class and binary classification are illustrated in Figure 4a and Figure 4b respectively. These results highlight FCN's robust ability to learn complex patterns and representations directly from raw sensor data, regardless of the signal source. The model's strong performance across all signal types demonstrates its effectiveness in capturing relevant stress-related features from both chest and wrist-worn devices, making it a good choice for stress detection applications. The particularly



(A) Three Class (B) Binary

FIGURE 4: Confusion matrices using FCN model

strong performance with combined signals suggests that FCN effectively leverages complementary information from multiple sources to enhance stress detection accuracy.

Of particular interest are the results obtained from wrist-worn devices, as these represent a more practical and less intrusive approach for real-world applications. The strong performance of FCN with wrist signals (86.42% accuracy in binary classification and 73.91% in three-class classification) is especially promising for future studies in controlled experiments focusing on affective and cognitive state detection in learning environments. While chest-worn devices showed slightly higher accuracy, the wrist-based measurements offer a good balance between performance and practicality, making them ideal for extended monitoring in experimental settings. These results demonstrate that wrist-worn devices, when coupled with advanced deep learning approaches like FCN, can effectively capture stress-related physiological patterns, providing a robust foundation for future research in affective and cognitive state detection.

Discussion

The analysis of the WESAD dataset revealed several important challenges and insights in stress detection using physiological signals. A primary challenge was the data imbalance, both in terms of signal sources and class distribution. The chest-worn device (RespiBAN) collected significantly more samples (4,255,300) compared to the wrist-worn Empatica E4 (194,528) for each subject. The dataset exhibited considerable label bias, with baseline samples outnumbering other classes and stress samples being approximately double the number of amused samples. To address this imbalance, SMOTE (Synthetic Minority Oversampling Technique) was employed in the analysis, generating synthetic samples for minority classes while avoiding the overfitting issues commonly associated with random oversampling.

A significant technical challenge arose from the varying sampling frequencies across devices and modalities. The chest-worn device sampled all signals at 700 Hz, while the Empatica E4 used different sampling rates for different modalities: BVP (64 Hz), EDA (4 Hz), TEMP (4 Hz), and ACC (32 Hz). This disparity in sampling rates complicated the selection of appropriate window sizes for feature extraction, particularly when combining data from multiple modalities. The manual feature extraction process required high domain expertise, highlighting one of the key limitations of conventional machine learning approaches.

The deep learning-based FCN demonstrated superior performance compared to conventional machine learning models, particularly with chest signals, achieving 91.54% accuracy for binary classification and 80.23% accuracy for three-class classification. This superior performance can be due to two factors: the higher sampling rate of chest signals providing more information, and the FCN's ability to automatically learn relevant features from raw data, eliminating the need for manual feature engineering. The model's architecture, particularly the filter size in convolutional blocks, proved crucial for performance optimization. While the exclusion of ACC signals significantly impacted performance, TEMP signals had minimal effect, suggesting varying importance of different physiological signals in stress detection.

TABLE 3.4: Classification accuracies when each sensor data is excluded for wrist data (FCN)

Signals	Accuracy	F1-score	Precision	Recall
Without ACC	**0.68 ± 0.11**	**0.66**	**0.67**	**0.69**
Without BVP	0.70 ± 0.07	0.67	0.68	0.70
Without EDA	0.70 ± 0.09	0.68	0.71	0.70
Without TEMP	0.71 ± 0.10	0.69	0.71	0.71
All signals	0.73 ± 0.09	0.72	0.73	0.72

The performance difference between chest and wrist signals (chest: 91.54% vs. wrist: 86.42% accuracy in binary classification) highlights the impact of sampling frequency on model performance. These findings emphasize the importance of high-quality, high-frequency data in stress detection applications and suggest that future work should focus on developing methods to better utilize data from sensors with varying sampling characteristics.

To understand the relative importance of different physiological signals in stress detection, an ablation study was conducted by excluding individual sensor data from the wrist-worn device, as shown in Table 3.4. The results revealed that ACC

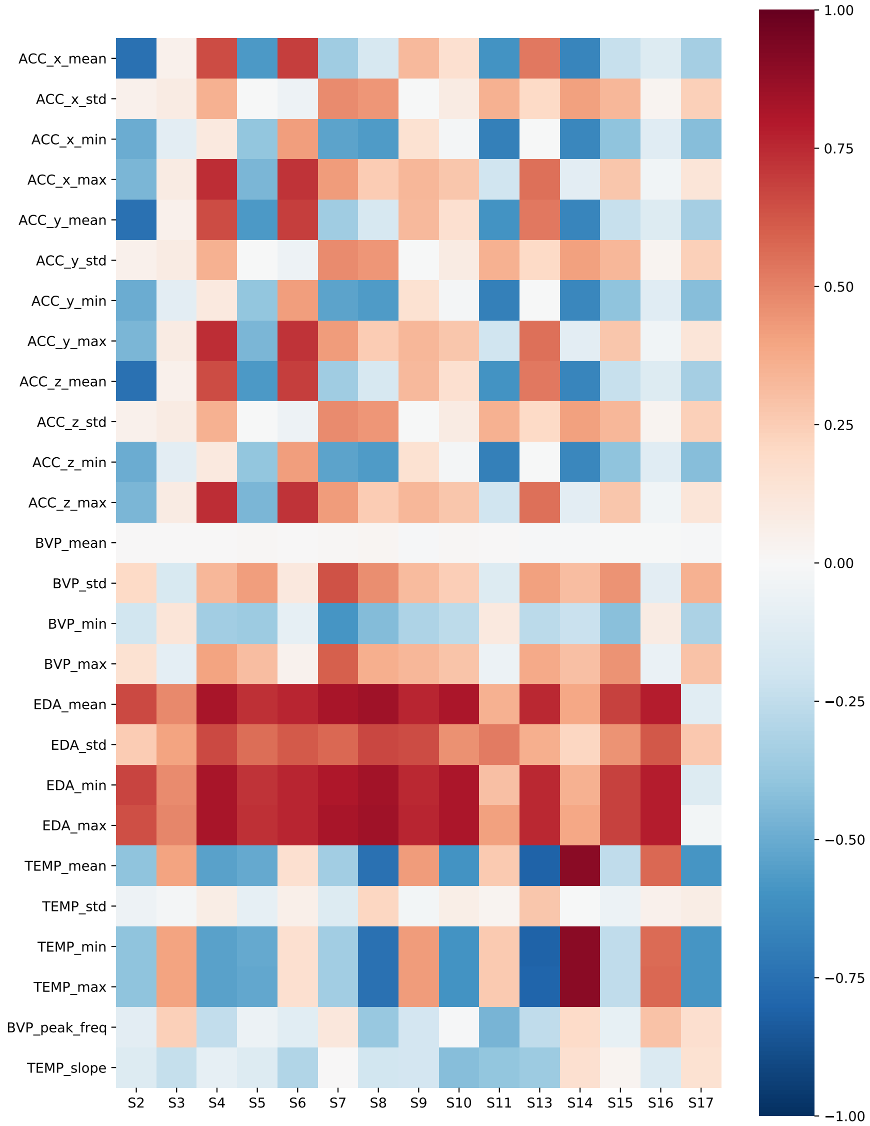


FIGURE 5: Pearson correlation of target variables to extracted features (wrist-data)

data had the most significant impact on model performance, with its exclusion leading to the largest drop in accuracy (68% ± 0.11) and F1-score (0.66) compared to using all signals (73% ± 0.09 accuracy, 0.72 F1-score). This suggests that patterns captured by the accelerometer provide crucial information for stress detection in the WESAD dataset. The exclusion of BVP and EDA signals resulted in a moderate performance decrease (70% accuracy for both), indicating their complementary role in stress detection. The removal of TEMP signals had the least impact on model performance (71% ± 0.10 accuracy, 0.69 F1-score), suggesting that temperature variations might be less indicative of stress states compared to other physiological signals. These findings provide valuable insights for sensor selection in wearable stress detection systems, highlighting the particular importance of accelerometer data while suggesting that temperature sensors might be optional in resource-constrained applications.

The analysis of individual participant responses revealed significant variations in both feature correlations and model performance. The Pearson correlation matrix

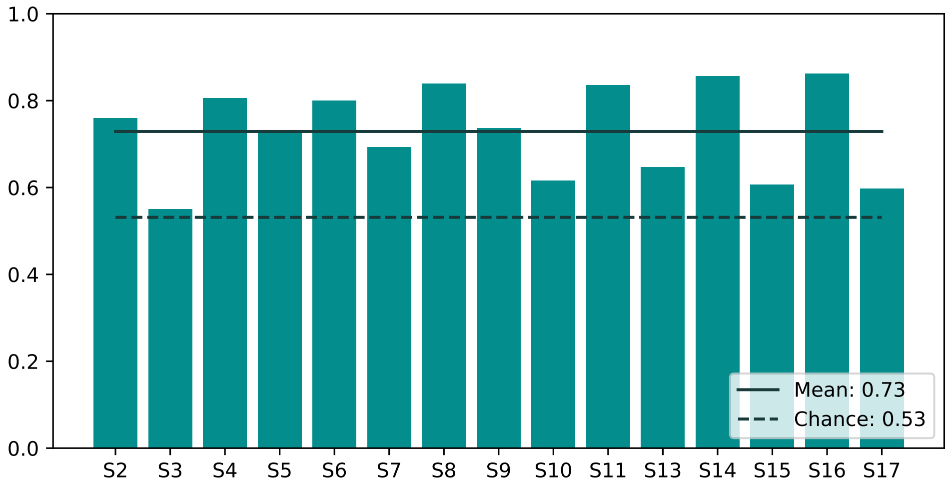


FIGURE 6: Accuracy per subject for three class classification using FCN and wrist data. Participant IDs are as they are in the dataset.

(Figure 5) shows that EDA based statistical features showed positive correlations with most participants, except for *S17*. The ACC features exhibited contrasting correlations across participants, with *S4* and *S6* showing positive correlations while *S2* and *S16* displayed negative correlations. This individual variability was further reflected in the participant-wise classification accuracies (Figure 6), where participants *S4*, *S8*, *S14*, and *S16* achieved higher accuracies when used as test sets, while others like *S3*, *S10*, *S13*, *S15*, and *S17* showed relatively lower accuracies. These variations in both feature correlations and classification performance can be attributed to the uniqueness of individual physiological responses and potentially different reactions to stress-inducing stimuli, highlighting the challenge of developing a generalized stress detection model.

3.1.2 Stress Detection using SWELL-KW Dataset

The SWELL Knowledge Work (SWELL-KW) dataset is a result of conducting an experiment with 25 participants doing typical office work like making presentations, writing reports, reading e-mails, and searching information. The data recorded during the experiment included facial expression, computer logging, body postures and three physiological signals based on ECG and skin conductance with the final goal of recognizing stress during working conditions [103].

The three different working conditions were :

- **Neutral "no-stress"**: The participants were allowed to work on tasks without any time limits and they are unaware of the maximum duration of the tasks.
- **Stressor "Time pressure"**: The participants had to finish the task within 2/3 time, that they did the task in neutral condition.

- **Stressor "Interruptions"**: The participants were interrupted during the tasks by sending e-mails, among which some were relevant for the task and some were not.

The participants performed knowledge worker tasks on a desktop computer in a controlled lab setting consisting of 6 selected topics on which people with various backgrounds could work. Participants completed a series of self-report questionnaires to collect ground truth data on their subjective experiences after each task block. The NASA Task Load Index [68] assessed task load, including mental and physical demand, temporal demand, effort, performance, and frustration. The Rating Scale Mental Effort evaluated mental effort, while the Self-Assessment Manikin Scale measured emotional responses in terms of valence, arousal, and dominance. The perceived stress was reported on a visual analog scale from 'not stressed' to 'very stressed' [103]

Methodology

While the primary focus is on deep learning approaches for sensor-driven analytics, the SWELL-KW dataset analysis was included to validate stress detection in a different context - specifically knowledge work settings. Unlike the WESAD dataset, where raw sensor data was utilized for analysis, the SWELL-KW dataset provides pre-computed features from four distinct categories of sensor data: physiological measurements, body posture information, computer interaction patterns, and facial expression data. This presents an opportunity to evaluate how conventional machine learning approaches perform with manually crafted features in a real-world office environment.

The physiological features include various metrics derived from ECG and skin conductance measurements, such as heart rate variability (HRV), skin conductance level (SCL), and root mean square of successive differences (RMSSD) between normal heartbeats. These pre-computed features, combined with behavioral indicators from body postures, computer interactions, and facial expressions, provide a comprehensive set of stress indicators in a knowledge work context. Table 3.5 shows the different features from each modality.

For the analysis, the stress detection was formulated in two ways. First, for three-class classification, the original condition labels were used: neutral, time pressure (encoded as stress-time), and interruption (encoded as stress-inter). Second, for binary classification, the two stressor conditions (time pressure and interruption) were combined into a single stress state (encoded as stress) and maintained the neutral condition as the non-stress state. Given the pre-computed nature of the features, a random forest classification was employed, which has shown robust performance in handling multi-modal feature sets.

Type	Features
Computer Interactions	Mouse (3)
	Keyboard (7)
	Applications (2)
Facial Expressions	Head Orientation (3)
	Facial Movements (10)
	Action Units (19)
	Emotion (8)
Body Postures	Distance (1)
	Joint angles (10)
	Bone Orientations (3×11)
	Other movements (44)
Physiology	Heart Rate (variability) (2)
	Skin Conductance (1)

TABLE 3.5: Features set from different modalities in the SWELL-KW dataset

Cross-validation Strategies: To evaluate the stress detection performance, two distinct cross-validation approaches were employed: user-independent and user-dependent approaches. In the user-independent approach, a Leave-One-Participant-Out (LOPO) cross-validation was implemented, where data from one participant was used as the test set while the model was trained on data from all other participants. This approach evaluates the model's ability to generalize across different individuals and assess its potential for real-world applications.

For the user-dependent approach, a personalized evaluation for each participant was performed. For each participant's data, k-fold cross-validation (k=5) was applied, splitting the participant data into train and test sets. The performance metrics were first calculated for each participant separately and then averaged across all participants to obtain the final performance measures. This approach provides insights into how well the model performs when trained and tested on individual-specific patterns, accounting for personal variations in stress responses. The combination of both strategies provides complementary insights: while user-independent validation demonstrates the model's generalizability across different individuals, user-dependent validation shows the potential performance benefits of personalization. This dual evaluation approach is particularly relevant in stress detection, as stress can vary significantly between individuals while maintaining certain common patterns across individuals.

Results

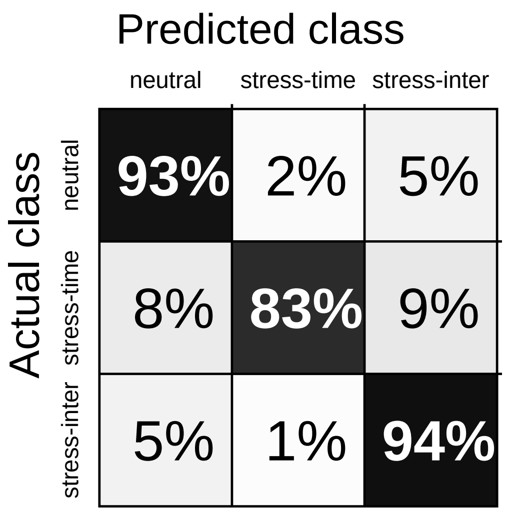
The classification results using random forest show significant variations in stress detection performance across different modalities and validation approaches, which can be observed from Table 3.6. In the user-independent approach (LOPO validation), the combination of modalities achieved the highest accuracy of 45.63% and 61.79% for three-class and binary classification, respectively. Among individual modalities, computer log data showed relatively better performance (43.28% for three-class), possibly due to its ability to capture work-related stress patterns through interaction behaviors. Physiological and facial modalities showed comparable performance (35.47% and 37.15% respectively, for three-class), while body posture features achieved slightly higher accuracy (39.82% for three-class).

	User-Independent		User-Dependent	
Modalities	Three-class	Binary	Three-class	Binary
Physiological	35.47	56.23	67.85	72.31
Body	39.82	58.64	80.29	85.76
Facial	37.15	56.92	72.43	79.58
Computer log	43.28	61.34	65.87	66.45
Combined	**45.63**	**61.79**	**91.24**	**92.68**

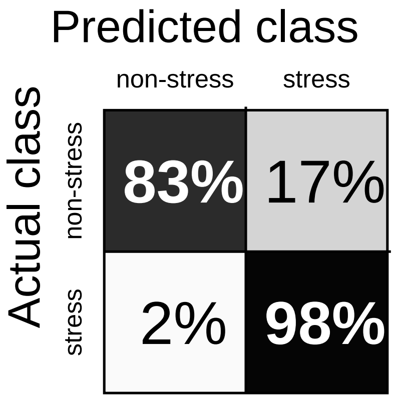
TABLE 3.6: Classification accuracy (%) using Random Forest for different modalities using user-independent and user-dependent approaches

The user-dependent validation revealed significant performance improvement across all modalities, highlighting the importance of personalization in stress detection systems. The combined modality approach achieved the best accuracy of 91.24% for three-class and 92.68% for binary classification, suggesting that integrating multiple data sources effectively captures individual-specific stress patterns. Body posture features exhibited good performance in the user-dependent approach (80.29% for three-class), followed by facial expressions (72.43%) and physiological signals (67.85%). This significant improvement in user-dependent results indicates that stress levels vary considerably between individuals, and personalized models can better capture these unique patterns.

The better performance of the combined modality approach in both validation scenarios highlights the complementary nature of different stress indicators. While computer interactions directly reflect work-related behaviors, physiological signals capture nervous system responses to stress. Facial expressions and body postures provide additional behavioral markers that might not be captured by other modalities. The performance gap between user-independent and user-dependent results (approximately 45% difference for combined modalities in three-class classification) emphasizes the challenge of developing generalized stress detection systems and suggests that hybrid approaches incorporating both personal and general patterns might be more effective for real-world applications.



(A) Three Class



(B) Binary

FIGURE 7: Confusion matrices using Random Forest

Discussion

The outcomes from multi-modal stress detection using the SWELL-KW dataset revealed several important insights into the effectiveness of different signal sources and validation approaches. The findings demonstrate both the potential and challenges in developing robust stress detection systems for workplace environments.

The better performance of the combined modality approach (91.24% for three-class classification in user-dependent scenarios) shows that stress affects both our behavior and body in different ways. Each modality contributes unique information: physiological signals provide objective measures of nervous system responses, facial expressions capture emotional states, body postures reflect physical variations of stress, and computer interaction patterns indicate changes in work behavior under stress. The effect of combining these modalities suggests that stress detection systems should ideally incorporate multiple modalities to capture the complete picture of an individual's stress state.

The significant performance difference between the user-dependent and user-independent approaches (approximately 45% for combined modalities) highlights a significant challenge in stress detection: the highly individualized nature of stress responses. The superior performance in user-dependent scenarios (ranging from 65.87% to 91.24% for three-class classification) indicates that stress variations are highly personal, with individuals exhibiting unique patterns across different modalities. This finding has important implications for the development of practical stress detection systems, suggesting that a hybrid approach combining personalized models with general patterns might be most effective.

These findings have several practical implications for workplace stress monitoring systems. The strong performance of user-dependent models suggests that stress detection systems should incorporate a calibration phase to learn individual stress patterns. The effectiveness of the combined modality approach indicates that workplace stress monitoring systems should utilize multiple data sources when possible, though the cost and complexity of implementation must be considered. Finally, the moderate performance in user-independent scenarios suggests that generic stress detection models might be suitable for initial deployment but should be adapted to individual users over time.

3.1.3 Limitations and Future Work

While this study provides valuable insights into stress detection using publicly available datasets, several limitations must be acknowledged. First, the datasets used, WESAD and SWELL-KW, were collected in controlled environments, which may not fully represent the complexity and variability of real-world stress scenarios. This controlled setting limits the generalizability of the models to more dynamic and unpredictable environments. Second, the sampling rates and sensor modalities differ between the datasets, which introduces challenges in standardizing feature extraction and model evaluation. Additionally, the reliance on high-frequency physiological signals, such as ECG and EDA, may not be practical for all real-world applications due to the cost and complexity of wearable devices.

Future work should focus on addressing these limitations by exploring stress detection in more naturalistic and diverse environments. Developing datasets that capture real-world stress responses across a broader population would enhance the generalizability of the models. The research into lightweight and cost-effective wearable devices that can capture multimodal data with lower sampling rates is essential for practical applications. Another promising direction is the integration of transfer learning techniques to adapt models trained on controlled datasets to real-world scenarios. Finally, privacy-preserving methods for collecting and analyzing physiological data should be prioritized to ensure ethical deployment of stress detection systems.

3.1.4 Conclusion

This research provides a comprehensive overview of stress detection methods using multimodal physiological data from two publicly available datasets: WESAD and SWELL-KW. The WESAD dataset, collected from 15 participants using a chest-worn device and Empatica E4, includes physiological signals such as ECG, EDA, EMG, RESP, TEMP, and ACC. The SWELL-KW dataset, focusing on HRV features, was also analyzed using machine learning classifiers, further demonstrating the effectiveness of modality-specific feature extraction. The WESAD dataset demonstrated the effectiveness of deep learning models, particularly the FCN, which outperformed traditional machine learning methods by automatically extracting features from raw data. The SWELL-KW dataset emphasized the importance of combining multiple modalities, with the combined approach achieving the highest accuracy in user-dependent scenarios. SWELL-KW revealed the critical role of personalization, as user-dependent models significantly outperformed user-independent ones, underscoring the highly individualized nature of stress responses. These findings have broader implications for the development of stress detection systems by highlighting the potential of integrating multimodal data to capture the complex nature of stress and the user-dependent models suggest that personalized approaches could significantly enhance the effectiveness of stress detection in real-world applications.

3.2 Stress Detection in Controlled Environment

This section and the related contents are based on our publication titled "*Enhancing Stress Detection for Students: Exploring the Impact of Fine-Tuning and User-Specific Data Calibration in Deep Learning*" [148]. In the previous section (Section 3.1), the analysis of stress detection using publicly available datasets was discussed. While these datasets provide valuable insights into stress detection, they often lack the specificity and control required to understand the variations of stress in specific populations. To address this limitation, a study was conducted in a controlled environment to detect the varying levels of stress in users. The study focused on students and aimed to investigate how they respond to different levels of stress [148].

The Empatica E4 wristband was used as a data collection tool to measure the stress levels experienced by students while they completed a series of tasks at different difficulty levels. The primary goal of the research was to assess and analyze stress in students, exploring how they respond to varying levels of stress and whether there were significant differences in their physiological reactions to these different stress levels. To achieve this, an experiment was designed and conducted involving 25 participants, who were subjected to three distinct stress conditions: 'easy', 'medium', and 'hard'. By employing mental arithmetic tasks as stressors, the intention was to generate stress-inducing scenarios at various intensities. The research aimed to enhance the understanding of how students experience and react to stress while performing tasks. By examining their physiological responses under different stress levels, the objective was to reveal patterns and variations that could explain the impact of stress on students' well-being and performance [148].

The main contributions of this study are [148]:

- A comprehensive and well-designed experimental protocol was developed, featuring a well-established stressor and a variety of tasks and stimuli that effectively induced different levels of stress in participants.
- A thorough evaluation and comparison of various deep learning models for predicting stress levels was performed, showing that these models can be successfully trained to precisely predict stress levels using physiological data.
- A comparative analysis was performed between a general classification model and a user-adaptive model, which was implemented by fine-tuning or calibrating user-specific information.
- An application prototype was developed that interprets dynamic variations of stress levels in users, utilizing a stress meter and customized alert messages, serving as a tool to help users understand and monitor their stress levels and ensure timely support.

This study highlights the importance of understanding and monitoring stress levels in students. By utilizing a robust experimental protocol, analyzing physiological data with deep learning models, and developing an application for visualization and potential interventions, the research contributes to the advancement of stress detection methods and their practical implementation [148].

3.2.1 Data Collection

The data collection was aimed at recording the physiological data from the participants while performing the tasks at varying levels of stress. The data collection and related subsections are based on our work [148].

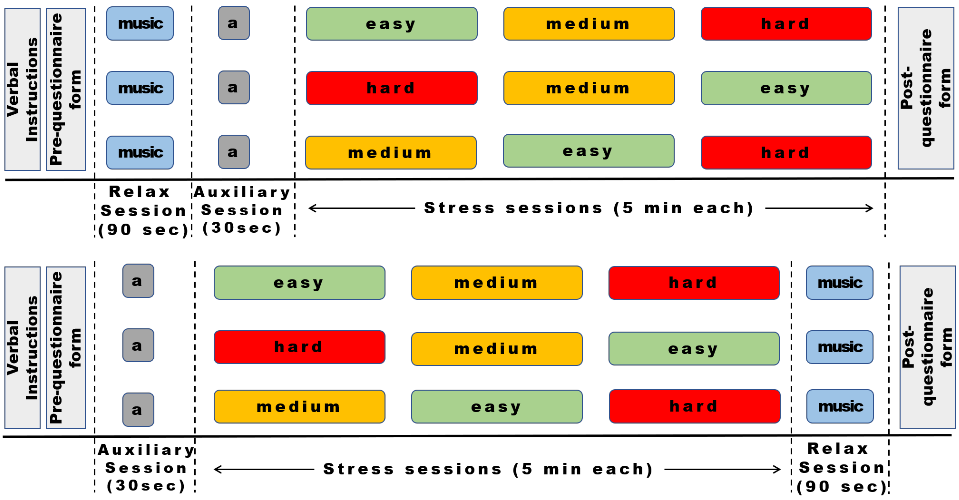
Participants

The study involved a total of 25 participants, comprising 11 female and 14 male participants, all of whom were pursuing their master's degrees at the university. The participants' ages ranged from 21 to 31 years. Prior to the experiment, an informed consent was obtained from each participant, ensuring that they were aware of the study's objectives, procedures, and potential risks. A detailed explanation of the experiment, including the tasks they would be required to perform, the data collection process, and the measures taken to ensure their confidentiality and anonymity was provided. The participants were also informed that they could withdraw from the study at any time without penalty or consequence. After obtaining consent, the experiment was conducted in a controlled environment to minimize external influences and ensure the collection of high-quality data. Figure 8a shows an overview of the experimental setup and how participant data is recorded [148].

Experimental Design

Before the experiment began, the procedure and apparatus were thoroughly explained to all participants. The participants who agreed and signed the consent form were allowed to participate. The Empatica E4 wristband was worn on the non-dominant hand, and participants were instructed to minimize hand movements to prevent introducing additional stress and variations. The experiment involved a series of tasks that lasted between 20 and 25 minutes. Participants were separated into two groups based on the sequence of relax and stress sessions: (1) relax followed by stress, and (2) stress followed by relax. These groups were then further divided into three subgroups according to stress levels: (1) easy-medium-hard, (2) medium-easy-hard, and (3) hard-medium-easy (Figure 8b). At the beginning and end of each relax and stress session, participants were instructed to press a button on the E4 wristband, indicated by a red blink. Clear instructions for the upcoming session appeared on the screen at its start. The different sessions involved in the experiment are explained below, as detailed in our previous work [148].

(A)



(B)

FIGURE 8: (a) An overview of the experimental setup. The participant is looking at numbers appearing on the screen in a series of mental arithmetic tasks. (b) The various sessions involved in the experimental setup

At the beginning and end of each relax and stress session, participants were instructed to press a button on the E4 wristband, indicated by a red blink. Clear instructions for the upcoming session appeared on the screen at its start. The different sessions involved in the experiment are explained below, as detailed in our previous work.

- **Relax:** Participants sat in front of a monitor displaying a pleasant image and listened to soothing music for 90 seconds.
- **Stress:** This session comprised three levels of arithmetic tasks, each lasting 5 minutes. A series of numbers was shown on the screen and participants were asked to memorize the series and mentally "Add/Subtract/Multiply" a given number to each number in the series. For example, if the series shown is (5, 3, 1) and if asked to add 1, then the correct answer will be (6, 4, 2). In the following fields, they were required to input the revised series. If the input

answer series is correct, an alert message "Correct" is popped on the screen and if not, an alert message "Incorrect" is popped. If they fail to input an answer, a "timeout" message appears on the screen as shown in Figure 9.

- **Easy:** Participants solved addition problems where they had to memorize 3 numbers in series and input 3 answers in the same order.
- **Medium:** Participants solved subtraction tasks where they had to memorize 4 numbers in series and input 4 revised answers in the same order. In addition to alert messages, participants heard a buzzer sound for correct, incorrect, and timeout results.
- **Hard:** Here participants solved multiplication tasks. The total numbers in the series were increased to 5. In addition to alert messages and buzzer sounds for correct, incorrect, and timeout results, participants heard a ticking clock sound when this session started.

• **Auxiliary session:** Before the stress session, a 30-second complementary task involved clicking on circles appearing on the screen. This session was not recorded using the wristband.

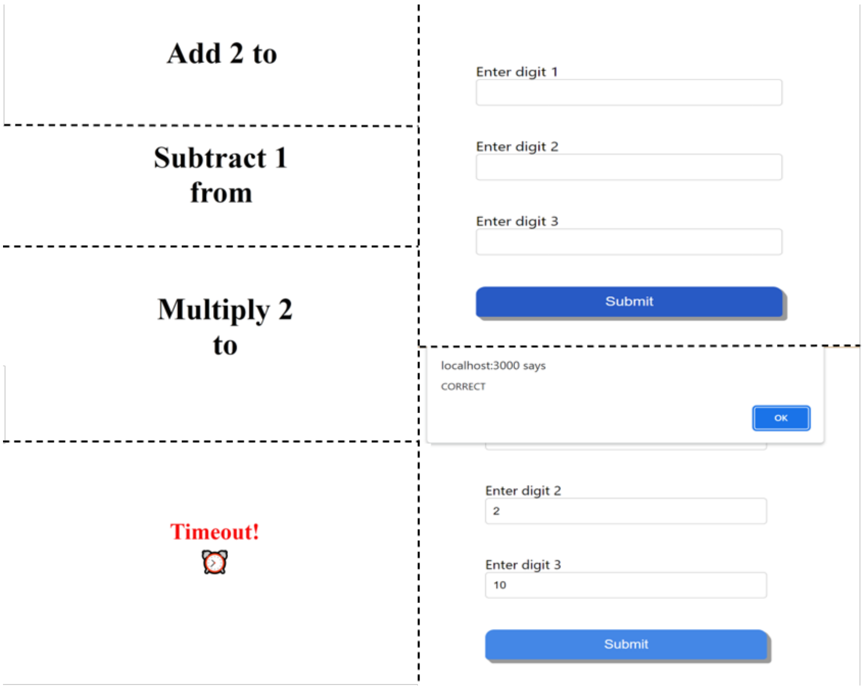


FIGURE 9: Screenshots illustrating the experimental sessions: easy, medium, and hard levels, along with a 'timeout' message (left column) and an example of the input field and 'correct' message (right column).

Participants took brief breaks between each session. In addition to the alert messages, buzzer and ticking sounds were employed to induce stress or anxiety. A questionnaire form was provided at both the beginning and end of the experiment, asking participants to answer several questions about themselves and the experimental process. The pre- and post-questionnaire forms utilized a Likert scale ranging from 1 to 7 (with 1 representing the least and 7 the highest) and included a drop-down menu for selecting options [148].

3.2.2 Methodology

To assess the varying stress responses in students, a controlled experiment was conducted using the Empatica E4 wristband to collect physiological data including Electrodermal Activity (EDA), Blood Volume Pulse (BVP), Skin Temperature (TEMP), Heart-Rate (HR) and three axis Accelerometer (ACC) data at 4Hz, 64Hz, 4Hz, 1Hz, and 32Hz, respectively. The system recorded physiological signals of the participant while performing a series of tasks at different stress levels. Figure 10 depicts the raw sensor signals recorded for a participant at the different sessions of the task. This methodology adopts an end-to-end approach, enabling the extraction of valuable insights directly from the raw sensor data without requiring manual feature extraction, as described in our work [148].

Data Pre-processing

The recorded EDA, BVP, TEMP, and ACC signals were preprocessed and segmented using a 30-second sliding window without overlap. All signals were standardized to have zero mean and unit variance before being fed into the models. To maintain consistency, some models received signals resampled to 4 Hz, while other models were trained using the signals at their original sampling frequencies. This methodology employs an end-to-end approach, enabling the extraction of valuable insights directly from the raw sensor data without manual feature extraction [148].

Model Architecture

The preprocessed signals were fed to multiple deep neural networks for predicting the varying stress levels in the users. The deep neural networks were employed which eliminates the need for manual feature engineering. Several architectures were explored, including convolutional neural networks (CNNs), such as fully convolutional networks (FCNs) and residual networks (ResNets), as well as recurrent neural networks (RNNs) like Long Short-Term Memory (LSTM) networks and transformer networks. The ResNet architecture (Figure 11a) featured a dense structure with multiple residual blocks, each containing three convolutional layers (64, 128, and 128 filters respectively) with varying kernel sizes and shortcut connections. Global average pooling downsampled feature maps, followed by fully connected layers for final classification. The FCN architecture consisted of three convolutional blocks per signal, each employing convolutional layers with ReLU activation, followed by global average pooling. The outputs were concatenated and fed into fully connected layers for classification [148].

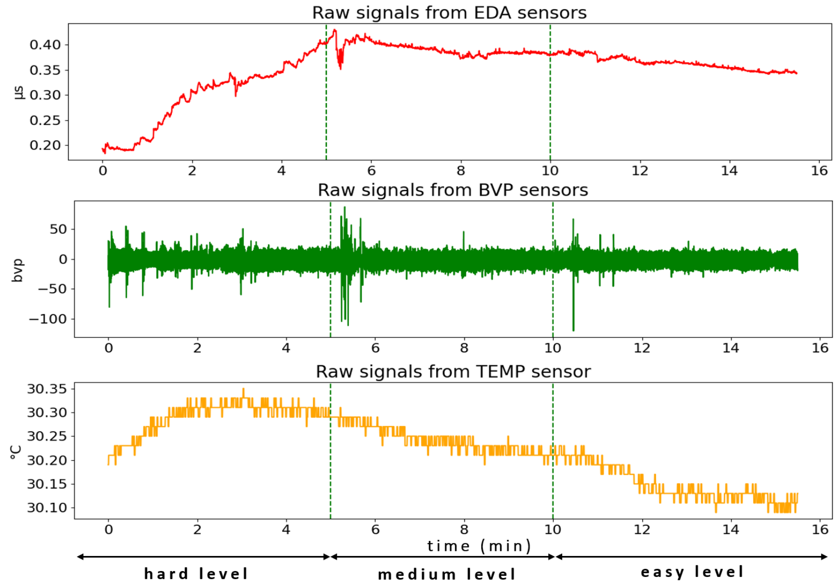
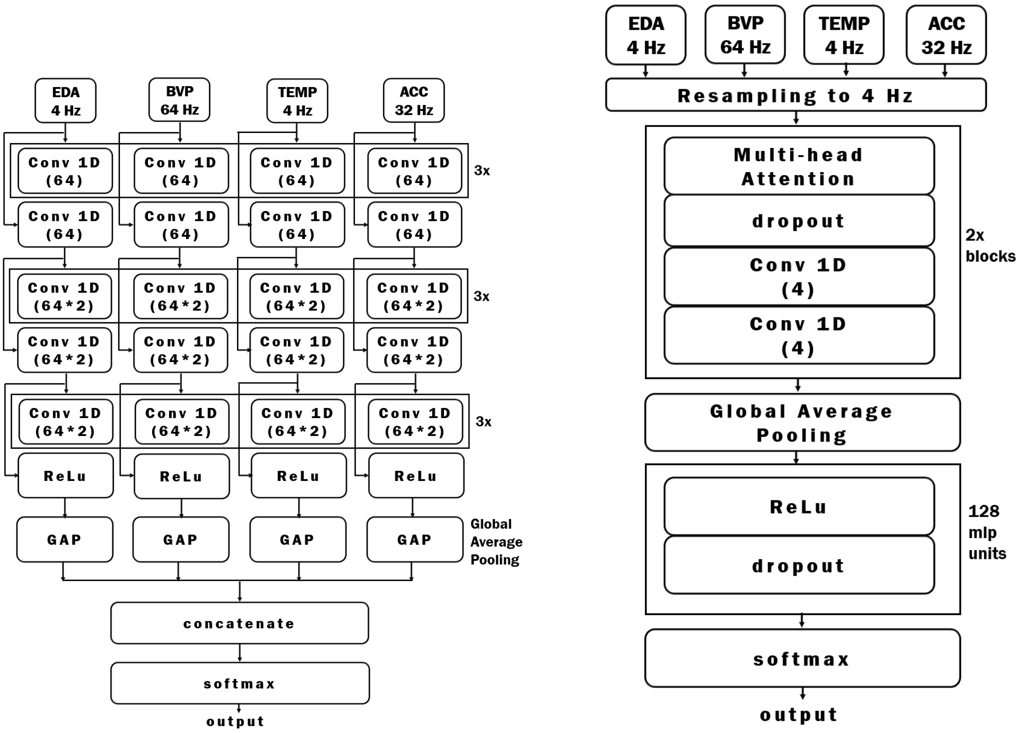


FIGURE 10: The raw physiological signals separated by different sessions of stress for a participant.

The transformer network (Figure 11b) processed the signals using two transformer blocks, each comprising a multi-head attention layer, a dropout layer, and two 1D convolutional layers. These blocks were followed by pooling and dense layers with ReLU and softmax activation. Finally, the LSTM architecture utilized a 1D convolutional layer (64 filters, kernel size 7), an LSTM layer, and dense layers with ReLU activations, followed by a dropout layer and a final dense layer. For the LSTM and transformer models, input signals were resampled to 4 Hz to ensure consistency [148].

Evaluation Protocols

Two cross-validation methods were employed to ensure robust and generalizable results: 5-fold cross-validation (KFold) and leave-one-participant-out (LOPO) cross-validation. KFold ensured balanced stress level distribution across the five subsets. LOPO trained models on data from all but one participant, then evaluated performance on the held-out participant's data. Considering the variability in individual reactions to stimuli, applying a generalized model to unseen participants presented a challenge. To overcome this, a fine-tuning (or calibration) approach was adopted, incorporating a small fraction (5%, 10%, and 20%) of the held-out participant's data into the training set to personalize the model. The remaining data (95%, 90%, and 80% respectively) was used for testing. For example, when evaluating participant p01, 10% of their data was added to the training data, and the remaining 90% was used for testing. This approach was followed to improve the model's ability to generalize to new, unseen participants by capturing individual variations in physiological responses. This systematic variation of training and testing data proportions



(A) (B)

FIGURE 11: (a) ResNet Architecture (b) Transformers Architecture

allowed for a controlled evaluation of model performance on unseen data [148].

3.2.3 Results

This study evaluated the performance of four deep learning models—FCN, ResNet, Transformer, and LSTM in classifying stress levels ('easy', 'medium', 'hard') using two cross-validation strategies: KFold and leave-one-participant-out (LOPO). The impact of fine-tuning on LOPO performance was also assessed. Results revealed significant differences in model performance and highlighted the benefits of incorporating user-specific data calibration. Table 3.7 summarizes the evaluation results for KFold and LOPO cross-validation, with and without fine-tuning, using different models. Across all models, K-Fold cross-validation achieved considerably higher accuracy and F1-scores than LOPO without fine-tuning. This highlights the challenge of generalizing models trained on one set of participants to completely unseen participants. The ResNet model outperformed other models, achieving 95.05% accuracy with KFold but only 35.52% with LOPO before fine-tuning. This suggests that ResNet, while performing exceptionally well within the training data distribution, struggles to generalize to new, unseen participants [148].

User-specific calibration/fine-tuning significantly improved LOPO performance for all models, demonstrating the effectiveness of incorporating user-specific data. The ResNet and LSTM models showed significant improvements. ResNet's accuracy increased from 35.52% to 85.12% with 20% fine-tuning data, while LSTM's accuracy

TABLE 3.7: Summary of classification results for three-class stress detection using the KFold and LOPO approach.

Model	Validation method	Acc	F1
FCN	KFold	85.05	0.848
	No ft	**41.16**	**0.483**
	5% ft	47.63	0.433
	10% ft	53.81	0.542
	20% ft	60.17	0.578
ResNet	KFold	**95.05**	**0.950**
	No ft	35.52	0.244
	5% ft	60.10	0.555
	10% ft	80.04	0.786
	20% ft	85.12	0.846
Transformers	KFold	59.12	0.572
	No ft	40.38	0.375
	5% ft	43.85	0.427
	10% ft	44.09	0.403
	20% ft	44.29	0.408
LSTM	KFold	78.21	0.781
	No ft	35.86	0.344
	5% ft	**71.09**	**0.709**
	10% ft	**84.12**	**0.839**
	20% ft	**91.17**	**0.911**

improved from 35.86% to 91.17% with the same amount of fine-tuning data. This suggests that these models benefit significantly from incorporating individual participant characteristics. In contrast, the Transformer model showed only marginal improvement with fine-tuning, indicating that this architecture may be less sensitive to individual variations. The FCN model showed a moderate improvement with fine-tuning, suggesting a moderate sensitivity to individual variations. Figure 12 visually depicts the distribution of predicted stress accuracies for each participant, comparing performance with and without fine-tuning. This figure allows for a detailed assessment of the impact of fine-tuning on both generalized and user-centric classification [148].

The confusion matrices for the LSTM model (Figure 13) provide a visual representation of the changes in classification performance with and without fine-tuning. These matrices illustrate how fine-tuning, even with a small percentage of user-specific data, enhances the model's ability to accurately predict the three stress classes. The results demonstrate the importance of considering both the model architecture

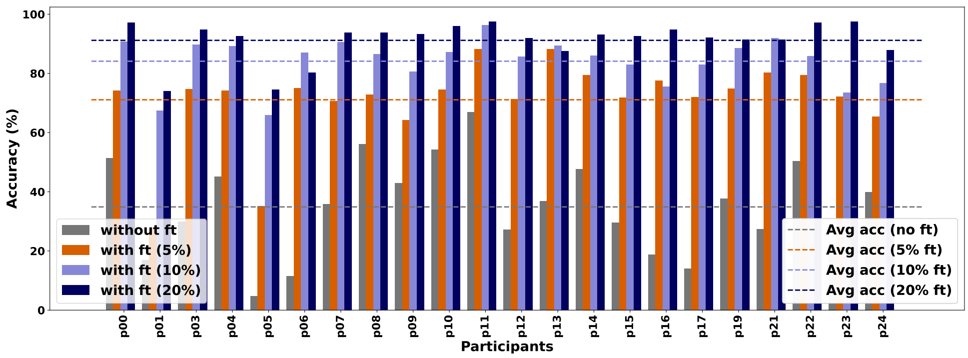
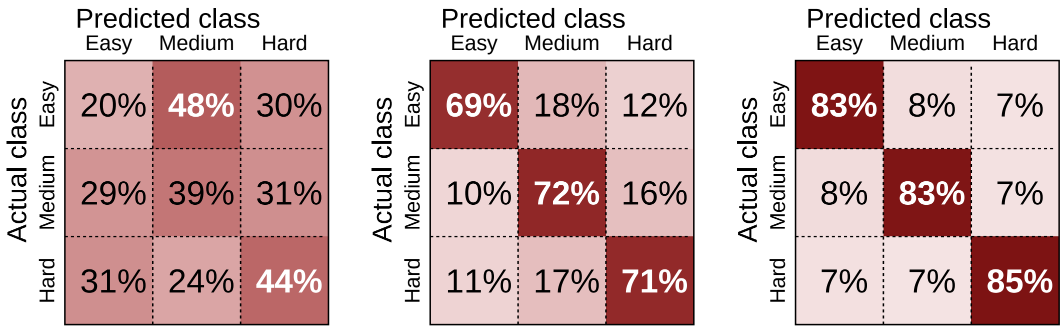


FIGURE 12: The prediction accuracies of LSTM using LOPO with and without fine-tuning (ft) for all participants.



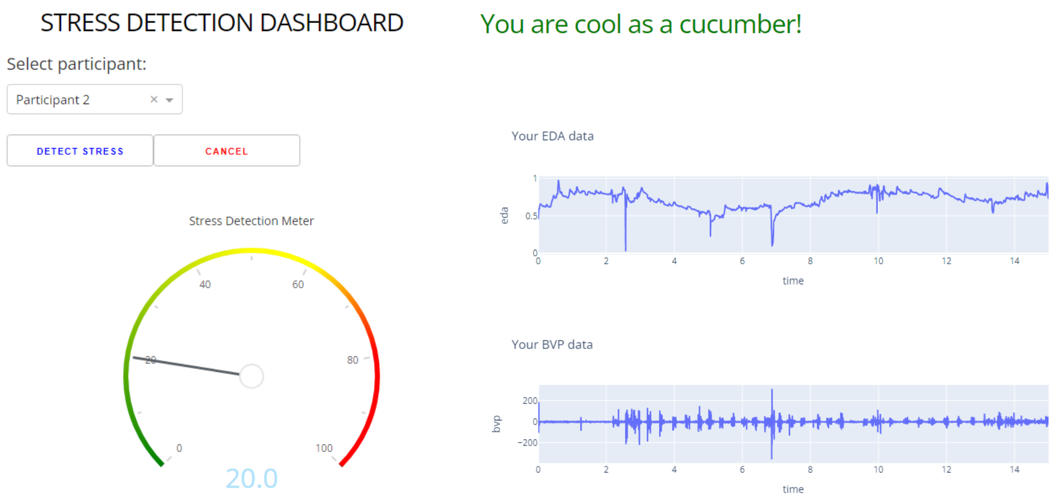
(A) No fine-tuning (B) With 5% fine-tuning (C) With 10% fine-tuning

FIGURE 13: Confusion matrices for LSTM model (a) without fine-tuning (b) with 5% fine-tuning and (c) 10% fine-tuning.

and the cross-validation strategy when developing stress classification models. Fine-tuning with user-specific data is crucial for improving the generalizability of these models, particularly for architectures like ResNet and LSTM that are sensitive to individual variations. The Transformer model, however, appears less susceptible to this issue. Further investigation into the reasons for these architectural differences could provide valuable insights into the design of more robust and generalizable stress classification systems [148].

3.2.4 Application

Beyond the implementation of the models, a prototype application was developed using Python Dash to visualize and represent the stress prediction results obtained from each participant. The application interface, shown in Figures 14a and 14b, was designed to offer a comprehensive overview of stress levels over time. Stress levels are evaluated and presented using a person-specific training method, which differs from the cross-validation techniques employed in the study's main evaluation. Each participant's data was split into training and testing sets using the KFold approach.



(A) (B)

FIGURE 14: (a) Stress-meter (b) Customized message with EDA and BVP signals

This ensures the stress visualization application effectively captures individual differences. This person-specific training improves the application's personalization, as a general model might not adequately account for individual variations, as detailed in our previous work [148].

The application enables users to select participants for analysis. Upon selection, dynamic changes in the stress meter and customized alert messages for each session (categorized as 'easy', 'medium', and 'hard') are displayed. The stress meter provides an overall indication of stress level, complemented by a feedback message offering additional insights. Electrodermal Activity (EDA) and Blood Volume Pulse (BVP) signals are also viewable alongside the stress levels, which provides a more comprehensive understanding of the relationship between stress and physiological responses. This application prototype functions as a potential intervention mechanism for students. By providing real-time insights into their mental state during various tasks, it enables students to effectively understand, manage, and address their stress levels. This tool could significantly contribute to student well-being and improve academic achievement by fostering self-awareness and proactive stress control [148].

3.2.5 Discussion

This research investigated the physiological responses of students to varying levels of stress, aiming to understand the impact of stress on well-being and performance as discussed in our previous work [148]. An experiment using mental arithmetic tasks as a primary stressor was designed. To induce varying stress levels, feedback, buzzers, timers, and clock-ticking sounds were incorporated. This controlled design, with predefined ground truths, minimized reliance on subjective questionnaire responses, promoting objective data assessment. The resulting multilevel classification system offers a more nuanced understanding of stress levels.

The deep learning models used in this study avoided the need for pre-processing or feature extraction from raw signals. However, accurately classifying three distinct stress levels from the Empatica E4 wristband's five sensory data streams presented a challenge. Due to the close relationship between heart rate (HR) and blood volume pulse (BVP) data (sampled at 64 Hz), HR data was excluded from the analysis. To ensure compatibility with the Transformer and LSTM models, which require uniformly shaped inputs, raw data was downsampled to 4 Hz. This downsampling, unfortunately, led to some data loss from the BVP and ACC sensors. In contrast, the FCN and ResNet models utilized all raw data at their original sampling frequencies. Despite these technical considerations, all models demonstrated reasonable performance, particularly with fine-tuning, showcasing the efficacy of deep learning for stress classification using E4 wristband data [148].

Initial findings from a participant-independent approach yielded unsatisfactory results, indicating substantial individual differences in how stress is experienced. However, classification performance improved considerably when models were fine-tuned using a small fraction (5%) of person-specific test data. This outcome underscores the critical role of fine-tuning for achieving better generalization and emphasizes the unique nature of stress responses across individuals. Personalized data calibration enables the models to adapt to these individual variations, resulting in more precise stress classification. The experimental setup involved four distinct sessions, comprising three stress levels and a 'relax' session. The 'relax' session was included with the specific intent of dividing participants into two groups: one group experiencing this session at the beginning of the experiment and the other group at the end. This design aimed to determine whether the timing of the 'relax' session influenced the levels of stress experienced by the two participant groups. Analysis of the outcomes revealed no significant differences in stress levels between these two groups. A substantial number of participants successfully completed their sessions, including the 'relax' component. A notable portion of these individuals faced challenges in correctly ending the E4 recording after the 'relax' session concluded. This issue led to inaccuracies in labeling the 'relax' session data, finally resulting in its exclusion from the model training phase [148].

3.2.6 Limitations and Future Work

This study, while demonstrating promising results, is subject to several limitations that require consideration for future research. These limitations primarily relate to data acquisition, experimental design, and the generalizability of the findings. The limitations and future work are discussed in our work [148].

- **Downsampling:** Downsampling the raw data to 4 Hz for compatibility with certain models resulted in a loss of information from the BVP and ACC sensors, potentially affecting the accuracy of stress classification.

- **'Relax' Session Data:** Difficulties in accurately recording and labeling data from the 'relax' session led to its exclusion from model training, potentially biasing the results and limiting the analysis of the relaxation state.
- **Generalizability:** While fine-tuning improved performance, the generalizability of the models to populations beyond the student participants in this study remains uncertain. Further research is needed to validate the models' performance on diverse populations and in different contexts.
- **E4 Wristband Limitations:** The study relied on data from the Empatica E4 wristband, which may not capture the full spectrum of physiological responses associated with stress. The use of additional physiological sensors could provide a more comprehensive assessment.
- **Induced Stress:** The induced stress, while controlled, may not perfectly replicate real-world stress experiences. The generalizability of findings to naturally occurring stress situations requires further investigation.

Future work should address these limitations by exploring higher-resolution data to mitigate the effects of downsampling and incorporating additional physiological sensors to capture a wider range of stress responses. The study should be expanded to include diverse populations and real-world stress scenarios that will enhance the generalizability of the findings. The future work could use unobtrusive wearable devices with enhanced sensing capabilities that could provide more reliable and detailed physiological data for stress analysis [148].

3.2.7 Conclusion

This study focused on examining user stress levels during mental arithmetic tasks using physiological signals from an Empatica E4 wristband. We used deep learning methods to predict these stress levels. The research included a user study with 25 university students, aiming to identify different stress levels labeled as 'easy', 'medium', and 'hard'. Among the various stress detection models proposed, both ResNet and LSTM showed strong predictive performance. This was especially evident when using K-Fold and Leave-One-Participant-Out (LOPO) cross-validation techniques, along with a fine-tuning method that included 5%, 10%, and 20% of the test data to improve personalization. With K-Fold, ResNet achieved a classification accuracy of 95.05% for the three classes, while LSTM reached 78.21%. With 20% fine-tuning, ResNet obtained an accuracy rate of 85.12% and an F1-score of 0.846. In comparison, LSTM achieved a higher accuracy of 91.17% and an F1-score of 0.911. For all four models implemented, applying fine-tuning with the LOPO cross-validation technique using 5%, 10%, and 20% of the test data consistently outperformed baseline methods. Overall, the study's findings confirm the effectiveness of deep learning methods in predicting stress levels from physiological signals. The results highlight the superior performance of the ResNet and LSTM models and show the benefits of fine-tuning or calibration to enhance the accuracy of stress level classification within a general prediction framework. This research sheds light on the potential of deep learning models in stress analysis and provides useful insights for future investigations in this area [148].

3.3 Multimodal Interest Detection

This section is based on our published work titled "*Multimodal Assessment of Interest Levels in Reading: Integrating Eye-Tracking and Physiological Sensing*" [157]. The fundamental aspects of human drive, the mechanisms of motivation, and the very nature of interest have been subjects of extensive inquiry long before the advent of digital technology [157]. Despite its frequent use, 'interest' remains a complex concept to define singularly, leading to various interpretations depending on context and specific aspects. Across the fields of psychology, sociology, and education, interest is commonly understood as "a construct that characterizes a person's special relationship with an object (contents, topics, special subject, and object domain)" [105]. Numerous researchers link it to academic achievement or define it as a motivational variable that is the foundation of productive learning [67, 104, 153]. Interest is also examined in relation to memory, with information deemed interesting by an individual being retained more easily and for longer periods than uninteresting information [150].

Within the context of reading, interest holds particular significance, given that learning and education are largely facilitated through this activity. Gaining insight into what stimulates the desire to read can greatly assist in designing documents that promote more efficient human-document interaction. The cognitive processes occurring in a reader's mind are reflected in their eye movements, causing changes in pupil diameter, blinks, fixations, saccades, and regressions [149, 198]. Thus, eye-tracking emerges as a direct method for detecting interest during reading tasks [14, 86]. Heart rate behavior represents another physiological channel through which emotions, and thus interest, can be effectively measured. Employing a wristband as a detector offers the advantage of being highly unobtrusive, thereby introducing virtually no distraction during the reading task [60].

This research investigates the identification of varying degrees of interest using raw data from eye-tracking and Empatica E4 wristband recordings as presented in our work [157]. The study compares the performance of a deep learning approach with previously established results obtained using conventional machine learning methods [88, 90]. Deep learning models—specifically, a One-dimensional Convolutional Neural Network (1D CNN), a CNN-LSTM (Convolutional Neural Network – Long Short-Term Memory) [127], and a Fully Convolutional Network (FCN)—were employed to extract high-level features directly from the raw sensor data, eliminating the need for manual feature engineering. This contrasts with conventional machine learning models, such as Random Forest (RF) and Support Vector Machine (SVM), which require explicit feature extraction and domain expertise. The proposed deep learning methodology leverages the inherent capacity of these models to handle temporal dependencies within the data—a crucial aspect for physiological signal processing—and to learn features at multiple levels of abstraction. Figure 15 provides an overview of the proposed system and facilitates a comparison between

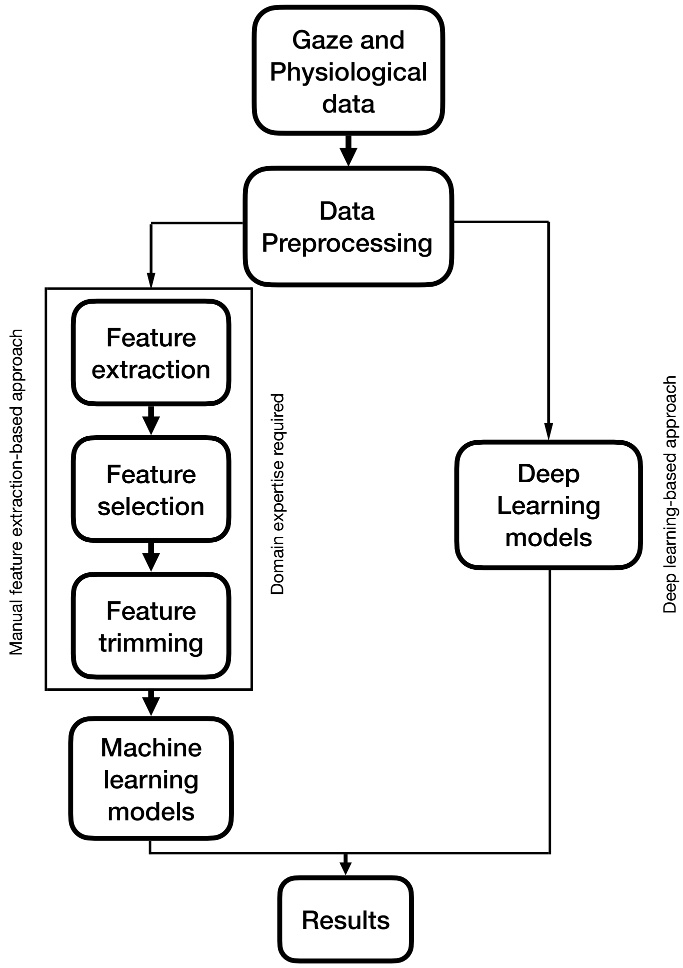


FIGURE 15: An overview of the manual feature extraction-based approach and deep learning-based approach

two distinct approaches: the conventional machine learning-based approach and the deep learning-based approach for interest detection. It illustrates the flow of data and processing steps involved in the proposed system, emphasizing the main stages and their connections. This comparative analysis allows for a direct assessment of the relative advantages of each approach within the context of this study. The research questions associated with this research are listed below [157]:

- **RQ1.** To what extent does a deep learning approach to interest prediction outperform a conventional machine learning approach that relies on manual feature extraction?
- **RQ2.** What is the impact of multimodal data fusion (combining eye-tracking and physiological data) on the accuracy and robustness of cognitive state (interest) prediction?
- **RQ3.** How does the predictive accuracy of personalized models for interest detection compare to that of generalized models trained on a diverse population?

To address these research questions, a comprehensive methodology was developed, incorporating data collection using eye-tracking and physiological sensors, the implementation of both deep learning and conventional machine learning models, and an evaluation framework based on established performance metrics.

3.3.1 Data Collection

The data collection process involved university students reading a series of newspaper articles while their gaze patterns were recorded using an eye-tracking device as presented in our work [88, 157]. The study aimed to analyze the relationship between gaze behavior and the participants' interest and comprehension levels.

Participants

The experiment involved thirteen university students (mean age: 25, standard deviation: 3; 6 male, 7 female), two of whom had prior experience with eye-tracking technology. Each participant was instructed to read 20 newspaper articles, with lengths ranging from 403 to 649 words (mean: 555, standard deviation: 70). Before starting the experiment, all participants were fully informed about the study's purpose, and their consent was obtained before proceeding [157].

(A) Document rated as very interesting   (B) Document rated as boring

FIGURE 16: Comparison of gaze patterns for documents rated as very interesting (a) and boring (b). (a) exhibits a higher number and longer duration of fixations, along with increased reading time, while (b) shows decreased number and shorter duration of fixations and reduced reading time.

Experimental Design

The newspaper articles were selected from BBC News, as they better captured the purpose of the experiment compared to other text sources. To ensure a diverse range of topics, articles were chosen from categories such as technology, politics, sports, and cooking. Figure 16a shows the gaze pattern of a document rated as very interesting by participant *P9*, while Figure 16b shows the gaze pattern of the same document rated as very boring by participant *P11*. After reading each article, participants completed a survey to assess their comprehension and interest levels. Table 3.8 presents the survey questions, which included subjective comprehension (Q1), interest level (Q2, serving as ground truth), and objective comprehension (Q3). Participants rated Q1 and Q2 on a scale of 1 to 4, with 1 representing the lowest rating and 4 the highest. Q3 was a multiple-choice question specifically related to the content of the article to evaluate objective comprehension [157].

TABLE 3.8: The survey used for the experimental design

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

3.3.2 Methodology

The system was engineered to capture both eye movements and physiological signals from the participants. Data recording was performed using an SMI REDn Scientific 60Hz remote eye tracker along with an Empatica E4 wristband. The gaze data obtained from the eye-tracker included a timestamp, the x and y coordinates for both left and right gazes (relative to the screen edges), and the pupil diameter of each eye. Physiological data collected by the E4 wristband sensor consisted of 3-axis Acceleration (ACC; 32 Hz), Blood Volume Pulse (BVP; 64 Hz), Electrodermal Activity (EDA; 4 Hz), Skin Temperature (TEMP; 4 Hz), and Heart Rate (HR; 1 Hz). Figure 17 illustrates the raw sensor signals recorded from a single participant, corresponding to their interest ratings for various documents. The EDA signal appeared to increase in response to *high-interest* ratings while decreasing for *low-interest* ratings, with the opposite trend observed for temperature. Data from two articles were excluded for all participants due to an error during collection. For the analysis, the preprocessed data was segmented using a sliding window algorithm, employing a window size of 10 seconds with a 50% overlap, as detailed in our previous work [157]. This subsection, along with its subsequent parts covering data preprocessing, feature extraction, and model development, is based on our research as detailed in [88, 90, 157].

Manual Feature Extraction-based Interest Detection

This section details the feature extraction process for both gaze and physiological data obtained from the Empatica E4 wristband. The features were subsequently concatenated to investigate the effects of sensor fusion.

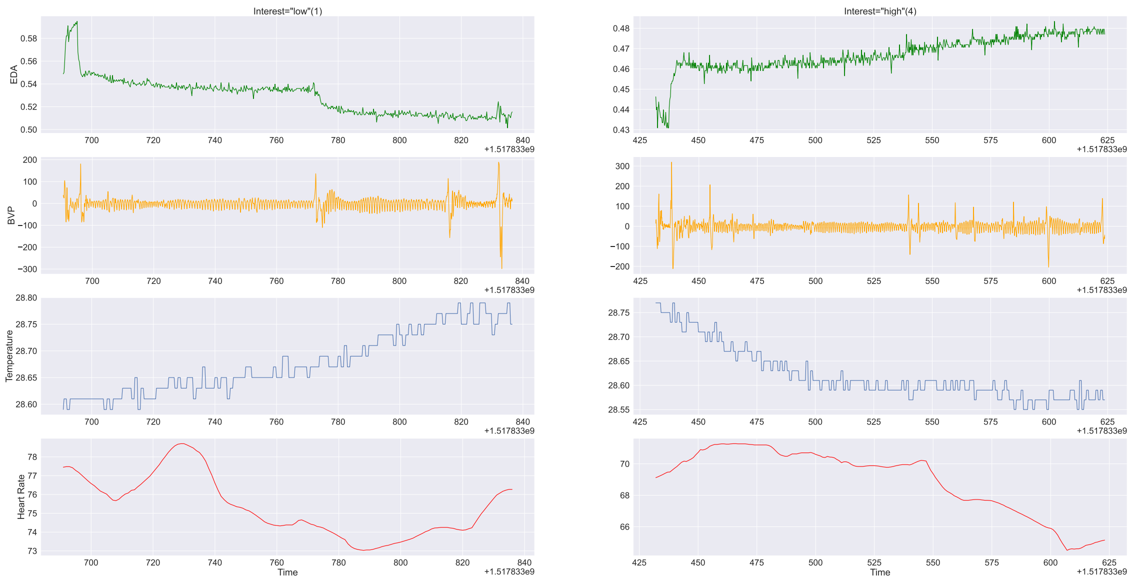


FIGURE 17: The raw physiological signals recorded from one participant, based on the interest responses for different documents.

Gaze Features: Raw gaze data underwent preprocessing to identify fixations and saccades. Seventeen features were then extracted from these fixations and saccades, including fixation duration, saccade length, saccade speed, regression length, regression speed, count of saccades, and count of regressions. Additionally, pupil diameter and blink frequency were extracted. Finally, mean and standard deviation were calculated for each of these features and used to train the model for interest classification [157].

Wristband Features: Following the preprocessing of the EDA signal from the E4 wristband, its phasic and tonic components were computed, as shown in Figure 18. In addition, the slope of the tonic component and the peak amplitudes for the phasic component were derived. From the BVP signals, heart rate (HR) and inter-beat intervals (IBI) were extracted. The mean and standard deviation were calculated for the BVP and HR data, and the difference in the mean and standard deviation of the HR amplitude between the task and baseline periods was estimated. Figure 19 displays the raw and cleaned BVP signal along with the extracted heart rate data. Similarly, the mean and standard deviation of the skin temperature (TEMP) were computed, and the slope of the TEMP signal was obtained. In total, 18 statistical features were calculated for the subsequent classification task [157].

To predict participants' interest levels, a Support Vector Machine (SVM) and a Random Forest classifier were employed. SVM operates by identifying an optimal hyperplane that maximizes the margin between different data classes, which is then used to classify new data points. SVM is particularly effective for high-dimensional data and can achieve high accuracy with complex datasets. Features derived from the eye tracker and those from the E4 wristband were used both separately and in combination for training these classifiers. The Random Forest classifier constructs

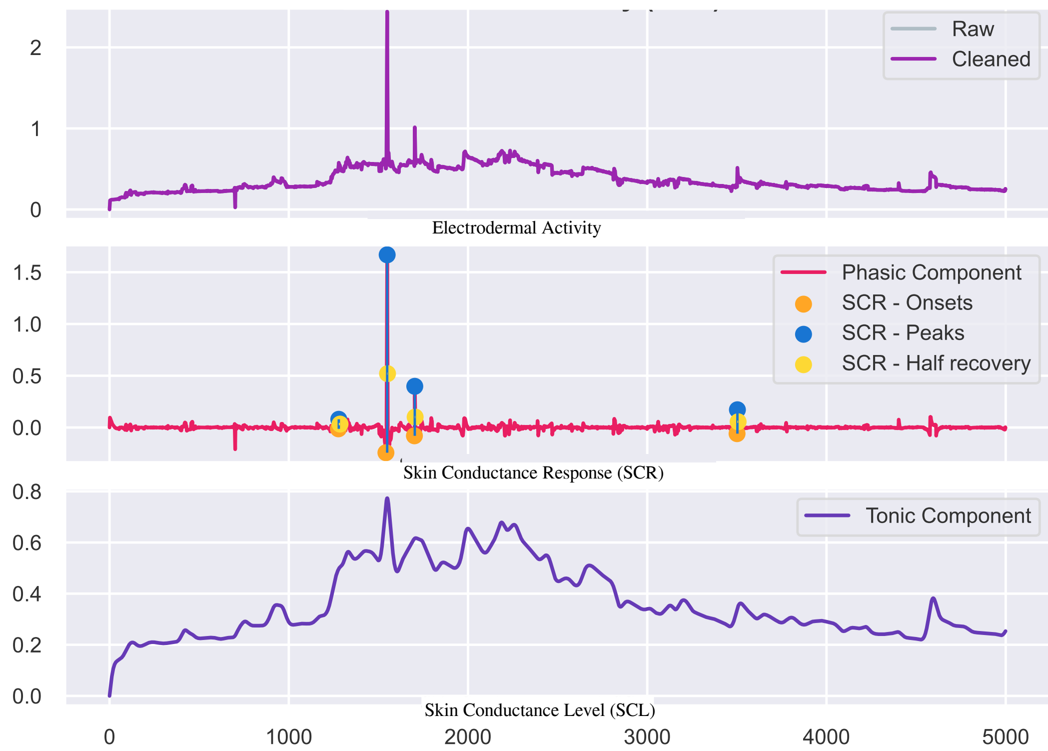


FIGURE 18: Preprocessed EDA signal showing phasic and tonic components.

multiple decision trees by selecting random subsets of training data, then combines the collective votes from these trees to determine the category of the test data [157].

The selection of appropriate hyperparameters (such as C, kernel, and gamma for SVM; and the number of estimators, maximum tree depth, maximum features, minimum samples required to split an internal node, and minimum samples required at a leaf node for Random Forest) is crucial but challenging. In this methodology, a 3-fold grid search cross-validation was utilized to optimize these hyperparameters. This technique explores a predefined set of parameters, identifying those that yield the highest validation score. A list of parameters was passed as arguments to the classifiers, and this process was repeated across different iterations of cross-validation to detect the most optimal parameters. Classification tasks were performed for both a four-class and a binary scenario, based on participant responses. In the four-class classification, labels ranged from '1' to '4', where '1' denoted the least interesting and '4' the most interesting. For the binary classification, labels '1' and '2' were merged and designated as 'not interested,' while labels '3' and '4' were combined and indicated as 'interested' [157].

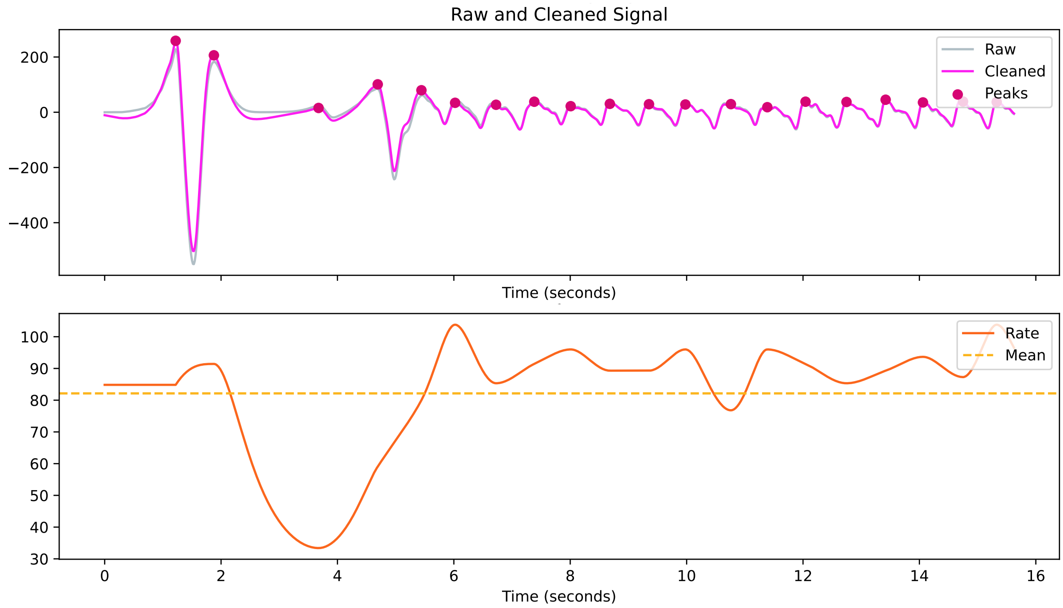


FIGURE 19: The peaks computed from the BVP signal and the extracted heart rate data

Deep Learning-based Interest Detection

Deep neural networks reduce the need for feature engineering, as they possess the capability to learn high-level features within their hidden layers. This simplification lowers the complexity of the workflow and the manual effort involved. It also increases the chances of discovering valuable information that might not be available to domain specialists. By using different layers, the structure of neural networks allows for a wide variety of deep learning architectures. For time series classification involving multiple physiological signals, a one-dimensional convolutional neural network has shown effectiveness in extracting relevant information directly from raw sensor signals, improving classification performance. Besides traditional machine learning classifiers, a deep learning model that combines 1D CNN, CNN-LSTM, and FCN architectures was used for interest detection. A key benefit of using deep learning models, compared to machine learning classifiers, is their ability to learn high-level features directly from raw sensor signals. This removes the need for manual feature extraction, which usually requires specialized domain knowledge. These deep learning models were trained with raw sensor input from both the eye tracker and physiological signals from the E4 wristband to predict interest [157].

FCN: This study utilizes a Fully Convolutional Network (FCN) architecture (identical to that presented in Section 3.1.1) for its ability to effectively process variable-length time series data without the need for padding or truncation. This is a significant advantage given the variability in the length of the physiological and gaze data recordings [157].

CNN-LSTM: CNN-LSTMs combine the strengths of Convolutional Neural Networks (CNNs), which excel at extracting local features from time series data, with Long Short-Term Memory (LSTM) networks, which are adept at capturing long-range temporal dependencies. The architecture typically consists of convolutional layers followed by one or more LSTM layers. The convolutional layers initially extract local features, which are then processed by the LSTM layers to capture global patterns and make the final classification. This combined approach allows CNN-LSTMs to effectively model both local and global patterns within the time series data. The architectures of the CNN-LSTM and a comparative 1D CNN model are shown in Figure 20 [157].

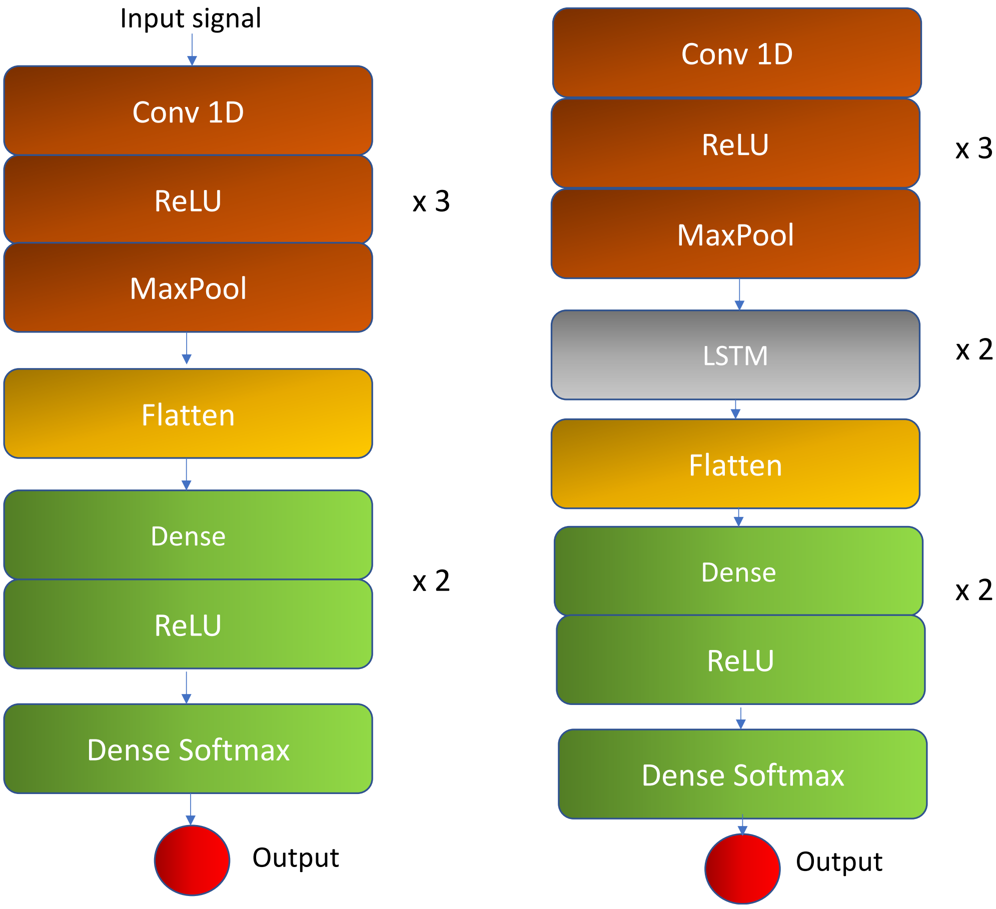


FIGURE 20: 1D CNN and CNN-LSTM architecture

Evaluation Protocols

Three distinct cross-validation techniques were employed to divide the data for training and testing. In the leave-one-participant-out cross-validation (LOPOCV) method, data from one participant served as the test set, while data from all other participants formed the training set for the model. This process was repeated for each participant, and the resulting accuracies were averaged to assess the model's overall performance. By evaluating the model on multiple test sets, each containing data from a different participant, a more thorough evaluation of its ability to generalize to unseen data was achieved. The averaged accuracies from these iterations provide a detailed performance evaluation that considers individual differences among participants. The leave-one-document-out cross-validation (LODOCV) technique involved excluding the data of a single document from the training set and using it as the test set. This process was iteratively applied for each document within the study, similar to the LOPOCV approach. Through the use of LODOCV, a deeper understanding could be gained regarding the model's generalization ability to unseen documents, and insights specific to the individual documents utilized in the study were obtained. The person-specific technique represented another strategy for data splitting, where data from a single participant (with each document from that participant used as a test) was divided for training and testing. The identical procedure was followed for other participants, and the average accuracy was derived from all participants. Given the unique nature of each person's engagement or interest, this strategy provides a more personalized approach to interest detection [157].

3.3.3 Results

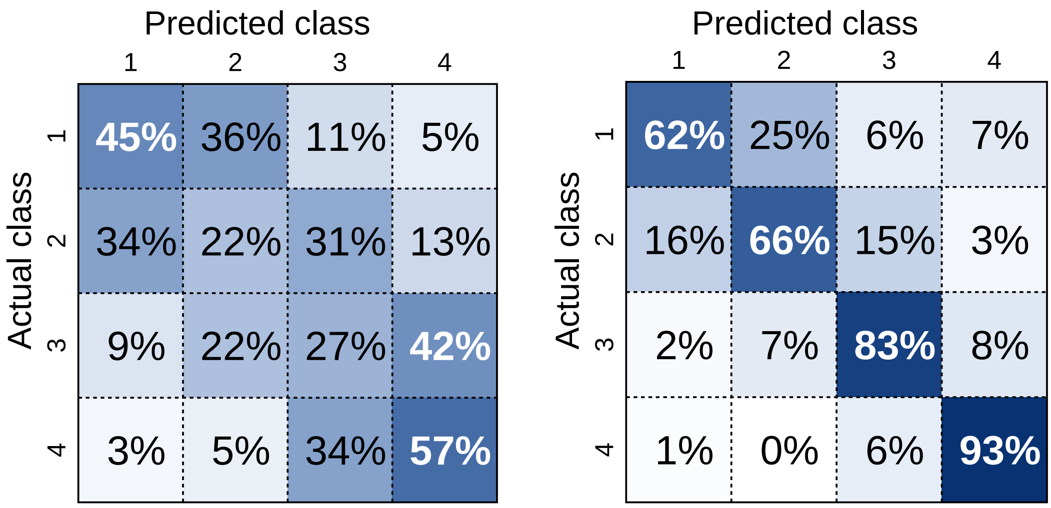
A baseline for comparison was established using a manual feature extraction approach. Random Forest and Support Vector Machine (SVM) models were trained to predict participant interest levels using features extracted from both gaze data and Empatica E4 sensor data. The SVM model consistently outperformed the Random Forest model in terms of accuracy and F1-score; therefore, only the SVM results are presented here. The interest levels were categorized into four classes: 1 (not at all interested), 2 (somewhat interested), 3 (mostly interested), and 4 (completely interested). Using Leave-One-Participant-Out Cross-Validation (LOPOCV), the SVM achieved an accuracy of 41.5% and an F1-score of 39.4%; Leave-One-Document-Out Cross-Validation (LODOCV) yielded an accuracy of 47.2% and an F1-score of 44.6% (Table 3.9)[157].

A binary classification task was also performed, grouping labels 1 and 2 as 'not interested' and labels 3 and 4 as 'interested'. In this binary classification, the SVM model, using combined E4 and eye movement features, achieved accuracies of 68% (LOPOCV) and 71% (LODOCV) (Table 3.10). For the deep learning methodology, several models were implemented: a 1D CNN utilizing gaze features, a CNN-LSTM model also employing gaze features, an FCN using physiological features from the E4 sensor, and a combined CNN-LSTM and FCN model that integrated both gaze and E4 data. Tables 3.9 and 3.10 display the four-class and binary classification results, respectively. The CNN-LSTM model, when using gaze features, significantly outperformed the SVM baseline, showing considerable improvements in accuracy for both binary and four-class classifications under both LOPOCV and LODOCV. Specifically, this CNN-LSTM model demonstrated a 14% accuracy enhancement for binary classification with LOPOCV and a 10% improvement with LODOCV[157].

TABLE 3.9: Summary of Four-Class Classification Results using machine learning and deep learning models for Leave One Participant Out (LOPO) and Leave One Document Out (LODO) cross-validation.

Models	**Features**	**Validation**	**Accuracy**	**F1-Score**
SVM	Gaze + E4	LOPO	41.5	39.4
		LODO	47.2	44.6
FCN	E4	LOPO	28.3	27.8
		LODO	33.9	32.0
CNN-LSTM	Gaze	LOPO	46.2	46.7
		LODO	**52.8**	**51.8**
CNN-LSTM + FCN	Gaze + E4	LOPO	42.7	40.2
		LODO	46.1	45.9

The integration of gaze and E4 features within the CNN-LSTM and FCN models did not produce the anticipated results across all evaluation metrics. Despite the expectation that combining gaze and E4 features would boost model performance, they did not exhibit the desired effect on metrics such as accuracy, F1-score, or other evaluation measures. This suggests that either the chosen features were not sufficiently informative or that the models were not effectively leveraging these particular feature combinations. A CNN model was also employed in the experiment; however, its prediction scores were lower than those of the CNN-LSTM model, leading to its exclusion from the table presenting the evaluation metrics [157].



(A) Participant-independent, acc.: 0.46

(B) Participant-dependent, acc.: 0.78

FIGURE 21: Confusion matrices for the four-class interest level classification where labels are 1: 'not at all interested (0%)´, 2: 'some of it (30%)´, 3: 'most of it (60%)´and 4: 'all of it (100%)'

TABLE 3.10: Summary of Binary Classification Results using machine learning and deep learning models for Leave One Participant Out (LOPO) and Leave One Document Out (LODO) cross-validation.

Models	Features	Validation	Accuracy	F1-Score
SVM	Gaze + E4	LOPO	68	65
		LODO	71	70
FCN	E4	LOPO	60.8	59.2
		LODO	61.4	60.1
CNN-LSTM	Gaze	**LOPO**	**82.3**	**81.7**
		LODO	80.5	80.2
CNN-LSTM + FCN	Gaze + E4	LOPO	71.5	70.2
		LODO	70.9	69.3

Beyond the two cross-validation techniques, which operate independently of specific persons and documents, a participant-dependent cross-validation method was also implemented. This was done recognizing that interest levels can be unique and vary based on the individual reading the text. In this person-dependent approach, each document from a given participant was held out for testing, and the average accuracy was then estimated. Figure 21 presents confusion matrices for comparison. As anticipated, a significant increase in accuracy was observed with this method, as the results were uniquely tailored to the specific individual and thus more personalized [157].

3.3.4 Discussion

This section addresses the research questions outlined in the introduction, analyzing the contributions, limitations, and challenges of this study. The discussion presented here is based on our published work [157].

Deep Learning vs. Manual Feature Extraction Approach

Previous studies [88, 90] employed manual feature extraction with SVM and Random Forest for interest prediction. While these methods offered initial insights, they were limited by their reliance on pre-selected features, which potentially failed to capture the complexities of reader interest. This reliance on set features can limit the ability to recognize all the complex and subtle patterns that affect reader engagement. The predictive accuracy might not fully reflect the dynamic nature of readers' preferences. SVM and Random Forest were chosen for analysis primarily because they are less susceptible to overfitting compared to other machine learning algorithms, making them suitable when data availability is limited. Random Forest also provides a measure of feature importance, which can be valuable for feature selection or engineering. In the current work, an interest detection method utilizing raw features with deep learning-based approaches was implemented, leading to an improvement in the efficiency of interest prediction. A comparison was conducted between the classification results achieved using an SVM model and those from deep learning models [157].

Based on the results obtained, it could be concluded that predictions were significantly enhanced by employing raw gaze features through the deep learning-based approach (Tables 3.9 and 3.10). The CNN and CNN-LSTM based models were able to surpass the accuracy achieved by the SVM model, which had used a combination of features from both gaze and E4 data. The primary objective of this study was to assess whether deep learning-based techniques could outperform manual feature extraction-based approaches in terms of prediction accuracy, and the findings indicated that deep learning models could be exceptionally beneficial in detecting cognitive states such as interest or engagement during a reading task. A key advantage of utilizing deep learning models for interest prediction is the elimination of the extensive effort involved in extracting the most relevant features for classification, a process that also demands high domain expertise. The work associated with identifying the most relevant features from physiological and eye gaze data could thus be eliminated with the deep learning-based architecture [157].

Impact of Multimodal Data Fusion

While combining gaze and E4 data improved accuracy for the manual feature extraction method, the deep learning approach showed different results. The raw gaze data alone gave a significant boost to the deep learning-based method. The CNN-LSTM model with the raw gaze features greatly improved classification accuracy. However, interestingly, combining the sensing modalities did not enhance performance compared to using only gaze features. The FCN model was trained with physiological signals from the E4 and then combined with the CNN-LSTM model to process the results by fusing both sensors. The prediction results from the sensor fusion with deep learning models showed only a slight improvement compared to the SVM classification accuracy. The raw gaze features with deep learning might be more accurate for predicting an individual's cognitive state. Although the performance of the deep learning models with combined sensor signals was similar to the manual feature extraction method, the raw gaze features from the eye tracker using deep learning models proved to be a better option for predicting interest [157].

Lin et al. explored how classification results were affected at both sensor and signal levels, noting that removing the EDA signal decreased accuracy and weighted F1 score for affect recognition, which offered a signal-level explanation [110]. Following their work, model performance was investigated by excluding each type of signal from the E4 wristband. Each modality was systematically left out to detect interest, aiming to understand the individual importance of each physiological signal in predicting user interest. Consistent with the results from Lin et al., Table 3.11

TABLE 3.11: The binary classification accuracy omitting each modality from Empatica E4. A decreased accuracy indicates the importance of the modality.

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

shows that accuracy was lowest when the EDA signal was removed, but omitting ACC and BVP had minimal to no impact on model performance [157].

Personalized vs. Generalized Models

A person's cognitive state can vary significantly depending on the context, and each individual approaches problems in a distinct manner. To investigate how person-specific predictions compare against generalized models, interest was estimated at a person-specific level by using data from a single participant for both training and testing. Each person's level of interest when reading a particular context can differ greatly, and what captures one individual's attention might not have the same effect on another [157].

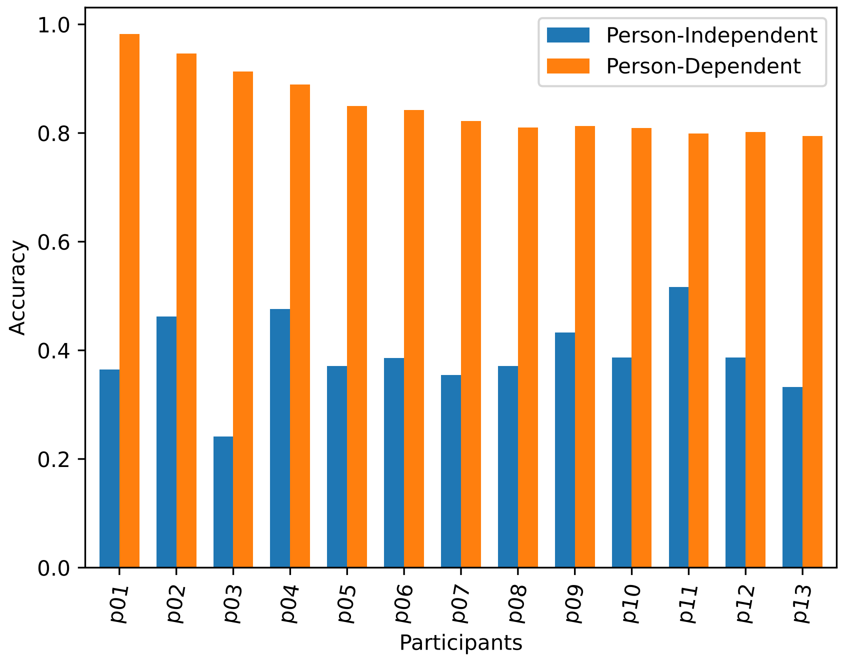


FIGURE 22: The plot showing the difference in accuracy per participant using a CNN-LSTM model with a person dependent and person independent approach

To account for this individual variability, a CNN-LSTM model, utilizing raw gaze features, was developed and deployed with a person-specific approach. This approach involved adapting the model to each individual's distinct characteristics and preferences. By integrating these unique factors, the model achieved substantially better performance in accurately identifying interest levels compared to generalized models. Indeed, the person-specific CNN-LSTM model showed an impressive performance gain of almost 50% per participant over other models. To demonstrate the effectiveness of this person-specific strategy for interest detection, Figure 22 illustrates the differences in accuracy observed across various participants. The visual representation provided by the plot highlights the individual accuracy scores, showing how the person-specific approach influenced the precision of interest detection. The variations in accuracy among participants further underscore the necessity of considering individual characteristics and preferences in this domain. The confusion matrices comparing participant-independent and participant-dependent methods (see Figure 21) offer further confirmation of a notable increase in accuracy when using a personalized approach, as the outcomes were uniquely tied to each specific person and thus more tailored [157].

Influence of Document Content and User-Specific Factors

The reading experiment featured a varied selection of materials spanning multiple categories, including technology, politics, sports, business, history, food, and more. Participants interacted with these materials, and their feedback differed greatly according to their personal interests. The purpose of the experiment was to gauge participants' subjective feelings of interest concerning the various documents provided. The choice of newspaper articles from BBC News was guided by their perceived capacity to align closely with the experiment's objectives. By selecting articles from this source, the goal was to guarantee a certain level of credibility, accuracy, and depth in the information presented. To better engage readers and sustain their interest throughout the experiment, the necessity of including a wide array of topics was acknowledged. Consequently, the search extended beyond a single platform, gathering articles from various sources and venues. This strategy enabled us to add a wider array of subjects that resonate with a diverse audience [157].

The intention behind including articles from a range of fields, including technology, politics, sports, cooking, and others, was to create a diverse dataset that would appeal to different reader interests. Moreover, efforts were made to reduce potential biases stemming from relying on a single news outlet by integrating content from various platforms. It is important to recognize that individual platforms tend to exhibit their unique editorial biases, writing styles, and viewpoints, which can impact the dataset. The objective was to assemble a more well-rounded and representative set of texts by collecting articles from multiple sources [157].

Figure 23 is referenced to evaluate the ratings given by participants for each document. Notably, document number 7 consistently received high ratings, suggesting it was deemed very interesting by most participants. In contrast, document number 16 received the lowest ratings and was viewed as the least engaging among the group. Further analysis revealed interesting insights concerning the document

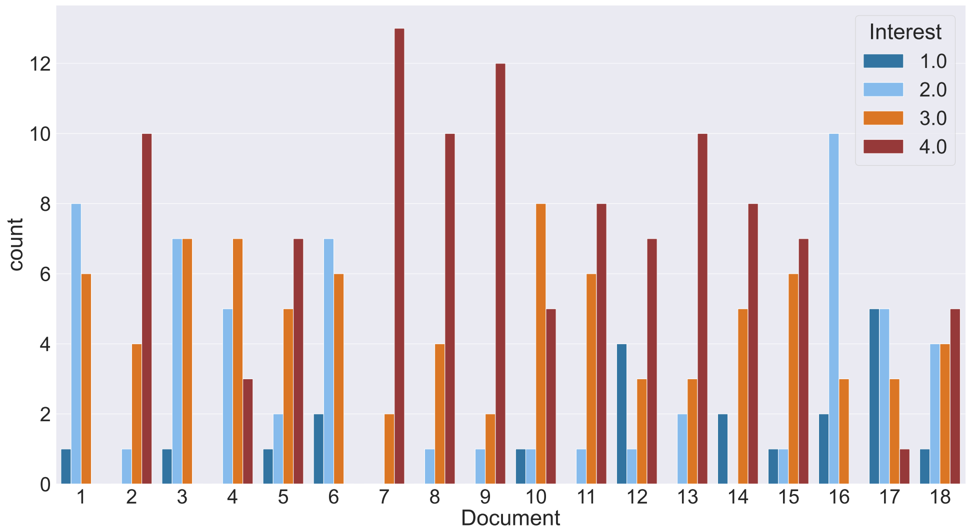
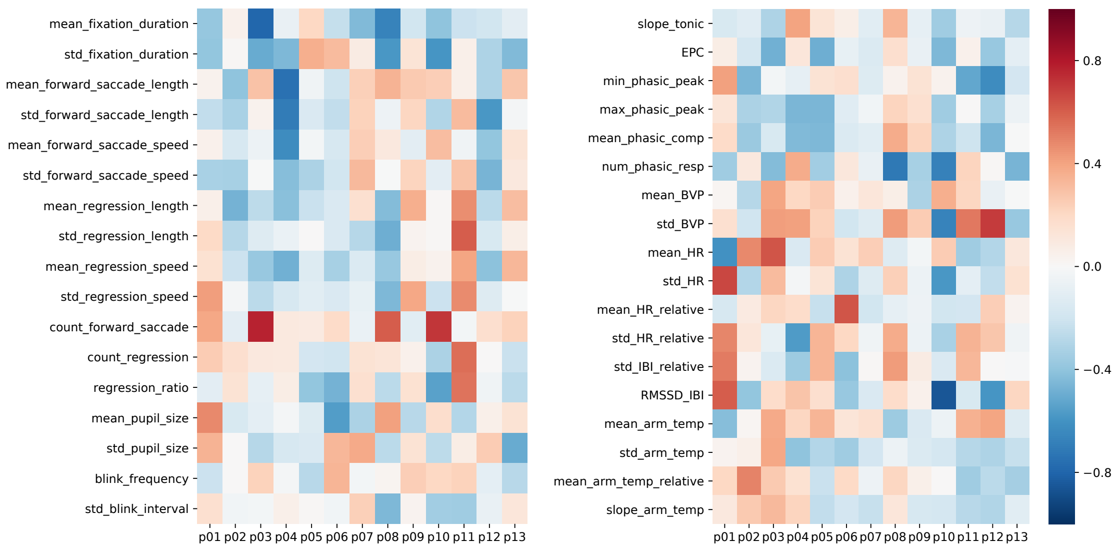


FIGURE 23: The interest ratings provided by the participants for all the documents used in the experiment

categories. It was noted that the document with the highest votes for being particularly interesting belonged to the *technology* category. This indicates that participants found the technological subject matter appealing and intriguing. Conversely, the document with the fewest votes, indicating the least interesting response, was categorized under *history*. This implies that the historical content did not resonate strongly with the participants. Given the varied reactions provoked by different documents, a thorough cross-validation process was executed using the Leave-One-Document-Out Cross-Validation (LODOCV) method. This approach allowed for a careful evaluation and validation of the models utilized in the study, ensuring the reliability and strength of the results [157].



(A) Eye-tracking (B) Physiological sensing

FIGURE 24: Pearson's correlations between interest and extracted features for each participant

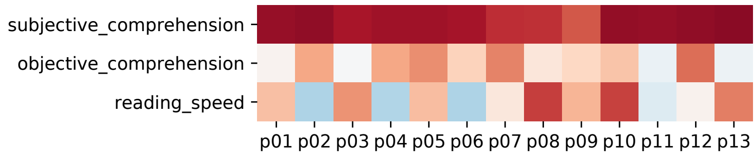


FIGURE 25: Pearson's correlations between interest and other surveys

The correlation plot illustrated in Figure 24 shows the relationship between handcrafted features and interest levels. It shows considerable fluctuations in the correlation between gaze features and interest level, depending on individual participants. This finding led to the interpretation that physiological predictions derived from gaze features depend more on the traits of individual users than on the documents themselves. In simpler terms, the connection between gaze features and interest is more related to the individual than the specific content of the documents. Figure 25 showcases the Pearson correlations between interest level and other measures not included in the proposed methods. Significantly, there is a strong correlation observed between understanding a document and interest in reading it. This conclusion aligns with the common understanding that one can only engage with interest if the text is comprehended. It underscores the importance of comprehension in the perception and engagement of interest during reading. Although incorporating user actions into the estimation task might have limited applicability, Figure 25 highlights that including user actions enhances the effectiveness of interest estimation. This suggests that additional user interactions, beyond gaze features, can lead to a more precise evaluation of interest levels [157].

3.3.5 Limitations and Future Work

This section, detailing the limitations of the current work and outlining directions for future research, is based on our published findings [157]. While deep learning-based approaches offer increased flexibility and effectiveness in predicting cognitive states, they are accompanied by certain limitations, including data scarcity, the high cost of training complex models, and significant computational power demands. For the present evaluation, data was gathered from 13 participants reading 18 distinct documents, which may be insufficient for training sophisticated deep learning models. Such models also necessitate substantial hardware resources for performing intricate mathematical calculations. The Empatica E4 recorded data at lower sampling frequencies, which could explain why the FCN model did not perform as effectively with the physiological data as initially anticipated. The initial expectation was that raw physiological features from the E4 wristband, when combined with a deep learning approach, would yield superior prediction results. However, the actual outcomes were comparable to those achieved using a manual feature extraction method. This suggests that the deep learning model's performance did not significantly benefit from the raw physiological features, indicating that the relationship between these features and interest was not effectively captured by the model. For future research, the intention is to employ more advanced sensors with higher sampling frequencies, which could enhance cognitive state detection using deep learning models [157].

The integration of gaze and E4 features with CNN-LSTM and FCN models did not produce the expected outcomes across all evaluation metrics. Further investigation and analysis are necessary to understand the reasons behind the suboptimal performance of these combined features with the CNN-LSTM and FCN architectures. The combination of gaze and E4 features was initially hypothesized to provide complementary information, with gaze features capturing visual attention and E4 features reflecting physiological signals. However, the unexpected results suggest that the information provided by these features might not be as complementary as first assumed. This could imply that the features are capturing redundant or irrelevant information, thereby hindering the models from effectively leveraging their combined strength. The quality and reliability of the gaze and E4 features are also crucial considerations. The combined feature approach might have been influenced by issues such as noise, outliers, or inconsistencies within the collected data. Similarly, it is possible that the chosen model architectures or hyperparameter settings were not optimal for effectively integrating the gaze and E4 features. Given these unexpected results, future directions involve exploring alternative feature combinations, investigating different model architectures, or considering additional relevant factors that may enhance the performance of interest detection models [157].

The selection of thirteen university students as participants in the experiment may indeed be considered a limited and potentially biased sample. While the study offers valuable insights, the small sample size restricts the generalizability of the findings to a broader population. As university students represent a specific demographic group with unique characteristics, the results may not be representative of the general population. Different age groups, educational backgrounds, and cultural factors could influence reading behaviors and eye-tracking patterns. Despite efforts to maintain a balanced gender distribution in the experiment (with six male and seven female participants), the unequal representation poses challenges in drawing conclusions uniformly applicable to both genders [157].

Numerous studies suggest that gender influences reading strategies, comprehension, and patterns of eye movement, emphasizing the importance of equitable gender representation in research. Moreover, the fact that only two out of the thirteen participants were familiar with eye-tracking technology presents an additional potential source of bias. Those who have used eye-tracking devices might exhibit different eye movement behaviors and reading patterns compared to those who have not. Such prior experience could impact the accuracy and reliability of the eye-tracking data, thus affecting the classifications derived from it. To address these limitations, future research should aim to recruit a more diverse participant pool, including individuals from varying ages, educational backgrounds, and gender identities, which would help improve the external validity and generalizability of the research findings.Another limitation of this study was the participants' inclination to be lenient in their ground truth responses. Although the selected articles for the experiment were carefully divided to evoke different levels of interest, a significant number of participants assigned 'highly interesting' ratings to the texts, leading to an imbalanced class distribution. For future studies, the intention is to create a system with established conditions aimed at minimizing this imbalance in class distributions and aiding in the collection of more reliable ground truth data [157].

3.3.6 Conclusion

The research aimed to evaluate the level of interest users exhibit while reading newspaper articles by integrating gaze tracking data with physiological responses. A multimodal methodology was utilized, merging gaze information with physiological indicators to offer a thorough and precise evaluation of interest levels. The study involved 13 university students engaging with 18 newspaper articles sourced from the BBC news database. Two strategies were implemented to predict the participants' levels of interest: one based on manual feature extraction and the other utilizing deep learning techniques. The results indicated that the deep learning-based method outperformed the manual feature extraction technique in predicting an individual's cognitive state, such as interest or engagement. A CNN-LSTM model that used raw gaze data achieved the highest accuracy of 52% for four-class classification and 82.3% for binary classification. A personalized approach tailored to individual participants was also used for more accurate predictions of interest levels, showing improved accuracy compared to the generalized method. To determine the importance of physiological signals in recognizing interest, each modality from the Empatica E4 wristband was excluded, revealing that the EDA signal plays a significant role in interest detection. These results enhance our understanding of interest in the reading context and suggest possible applications for personalized learning environments, improved understanding, and increased motivation in educational contexts. Future investigations in this domain could look into enhancing the measurement of interest and its influence on learning outcomes [157].

3.4 Deep Learning Based Emotion and Engagement Detection

This section is adapted from our published work titled *"Toward an Interactive Reading Experience: Deep Learning Insights and Visual Narratives of Engagement and Emotion"* [160] and *"Real-Time Feedback on Reader's Engagement and Emotion Estimated by Eye-Tracking and Physiological Sensing"* [138]. Recognizing the dynamics of engagement and emotion during reading activities is essential for enhancing comprehension and user experiences across various fields, including education, content production, and human-computer interaction [41]. The extent of engagement signifies how immersed and attentive the reader is toward the material, while emotions refer to the range of affective responses provoked by the text. A comprehensive understanding of these factors is vital in numerous domains, such as education, marketing, and user experience design [16, 91, 160].

In the context of e-learning, accurately gauging students' levels of engagement is critically important. Engagement serves as a significant measure of students' active participation and motivation in their learning journey, which has a direct effect on their academic performance and overall educational outcomes [7, 190]. To improve the effectiveness of e-learning platforms, it is crucial to utilize students' emotions as a means of evaluating their engagement accurately [49]. However, the current situation poses various challenges that hinder the integration of engagement and emotion analysis in e-learning. A major barrier is the lack of extensive recorded data that captures students' engagement and emotional reactions throughout the learning experience. This scarcity of data obstructs the creation of reliable machine learning models that can effectively predict engagement levels based on emotional indicators [160].

Ishimaru et al. proposed that aligning instructional content with the cognitive states of students may improve their learning and understanding, emphasizing the ability to predict a reader's internal states, such as their engagement level, to enhance the interactiveness of reading materials [86]. By understanding a reader's emotional and cognitive reactions, educators, and content developers can create a reading experience that is more engaging and customized. The emotional dimension of engagement is frequently neglected in current online education systems. Although traditional metrics provide information regarding student participation and completion rates, they do not capture the emotional experiences that motivate these behaviors. Relying solely on quantitative measures without exploring the emotional aspect may result in an incomplete view of students' true levels of engagement and could limit the opportunities for tailored learning experiences [160].

Current techniques for emotion detection are primarily focused on analyzing image data rather than physiological data [119, 120]. Although emotion detection based on images has shown potential across various applications, it may not adequately capture the complexities of emotional responses during the learning process.

FIGURE 26: The design overview

In contrast, analyzing physiological data offers significant opportunities to reveal subtle variations in emotions, such as fluctuations in heart rate, skin conductance, or facial expressions, which are closely associated with students' engagement and cognitive activities [143, 146, 160].

Beyond physiological signals, eye-tracking data is essential for identifying engagement and emotion during reading activities [176]. By monitoring the direction and movement of a reader's gaze as they engage with the text, gaze data offers important insights into their attention, focus, and cognitive processing. This gaze data can uncover emotional reactions, as a higher number of fixations on words or sentences with strong emotional weight may suggest increased emotional engagement. By integrating gaze data with machine learning and analytical methods, researchers can create models that effectively identify engagement and emotion, leading to a more comprehensive understanding of readers' cognitive and emotional conditions during reading [122, 160].

The integration of gaze data and physiological signals provides significant benefits for identifying emotion and engagement during reading activities. Brishtel et al. illustrated the promising capabilities of the EDA sensor for identifying mind wandering, and the combination of electrodermal activity with eye-tracking data greatly improved the precision of mind wandering classification [24]. Gaze information offers crucial insights into the reader's visual focus and attention toward particular parts of the text, whereas physiological indicators like heart rate and skin conductance provide objective metrics of emotional arousal and valence [115, 177]. By merging these modalities, a deeper understanding of the reader's cognitive and emotional conditions can be achieved, leading to enhanced accuracy and reliability in detecting emotion and engagement. This multimodal strategy allows for real-time evaluation, promoting personalized and adaptive reading experiences tailored to the reader's emotional reactions. Additionally, the unobtrusive nature of the data gathering process cultivates a natural reading setting, improving the generalizability and relevance of the results across various populations and reading scenarios [160].

The motivation for this research was to investigate the relationship between engagement and induced emotions in a learning environment for students. Previous studies have shown that materials that generate positive feelings, such as happiness or inspiration, tend to enhance engagement, while materials that provoke negative feelings, like sadness or boredom, often reduce engagement. Rosa et al. discovered a significant link between the experience of positive emotions, especially enjoyment, and the use of deep learning strategies among students [152]. This link was found to be closely related to students employing a dual approach that includes both mastery-focused and performance-focused goals in their learning endeavors. The study by Huang et al. strengthens the emphasis on positive emotions, confirming that feelings such as happiness and enjoyment lead to increased interactions between students and instructors [201]. Understanding the connection between engagement and emotional states could offer a more detailed perspective on user experiences, and by integrating these findings, content creators could refine their methods to not just grab attention but also elicit positive emotional reactions, nurturing a deeper and more impactful relationship with their audience [160].

In this context, our emphasis is on identifying engagement and emotion during reading tasks through a multimodal approach that combines gaze data and physiological responses from participants. The design overview is illustrated in Figure 26. For our research, we utilized various deep learning models to analyze the collected multimodal data. This allows us to offer a thorough comparison of the effectiveness of different models and to understand which ones are most capable of addressing this intricate task. An additional aim of our study is to assess the correlation between engagement and emotion during reading activities. Gaining insight into this relationship can provide valuable information regarding the dynamic interaction of cognitive and emotional processes in reading. By interpreting this connection, we may enhance the development of interactive reading systems and foster better reading engagement and comprehension strategies [138, 160].

The research questions of this research are the following :

- **RQ1.** To what extent do different deep learning architectures vary in their accuracy and generalizability when classifying user engagement and emotional states from multimodal data collected during reading tasks, considering both user-independent and user-dependent approaches?
- **RQ2.** What is the nature and strength of the correlation between engagement and emotion during reading tasks, as measured through multimodal data analysis?
- **RQ3.** How do individual differences in the relationship between engagement and emotion affect the performance of deep learning models in predicting these states during reading?

3.4.1 User Study and Data Collection

The data collection procedure was designed to record the participants' personal responses regarding their levels of engagement, valence, and arousal while they completed a reading task. The subjective evaluations obtained from the participants served as the ground truth labels for the analysis of the data. The user study including its subsequent subsections is based on our published research [138, 160].

Participants

The study involved the recruitment of 20 master's degree students from university, consisting of 11 males and 9 females. Their ages varied from 22 to 28 years, with an average of 25 years. All participants joined the experiment after giving informed consent, and they had the option to withdraw from the study at any time if they desired. Detailed information regarding data consent, the use of sensors, and the experiment sessions was provided to all participants prior to their involvement in the study. Data from three participants were eliminated due to data collection errors, which will be addressed later in the results section [160].

Experimental Design

The study featured a carefully designed reading task focused on comprehension, aimed at examining the diverse levels of engagement and emotional responses generated in participants. To accomplish this, we assembled a varied selection of 14 documents, each intentionally selected to provoke unique emotional reactions and engagement outcomes. Figure 27 depicts the experimental arrangement, highlighting a participant reading text from a screen equipped with an eye-tracker while wearing a wristband that records physiological data [160].

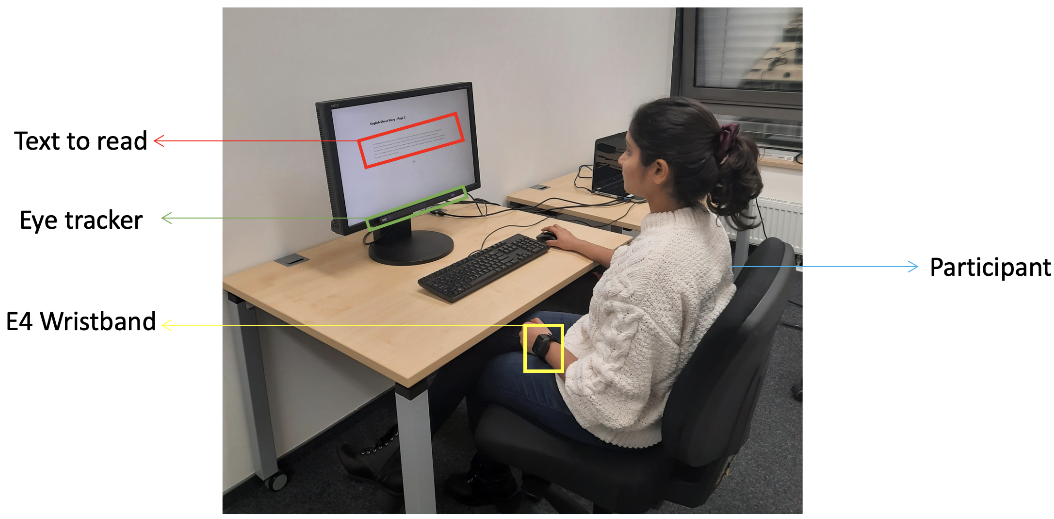


FIGURE 27: The experiment environment

Throughout the experiment, each document was presented to the participants one at a time. After reviewing each document, they were prompted to rate their engagement, valence, and arousal using a '1' to '5' scale. A score of '1' represented the lowest levels of engagement, valence, or arousal, while a score of '5' indicated the highest. Table 3.12 displays the self-report survey of questionnaires along with the corresponding response scale [160].

TABLE 3.12: Self-report survey for experimental design

No.	Question	Type	Scale
1.	Engagement while reading	Engagement	1-5 (1: lowest, 5: highest)
2.	Emotion type (positive/negative)	Valence	1-5 (1: lowest, 5: highest)
3.	Emotion intensity	Arousal	1-5 (1: lowest, 5: highest)

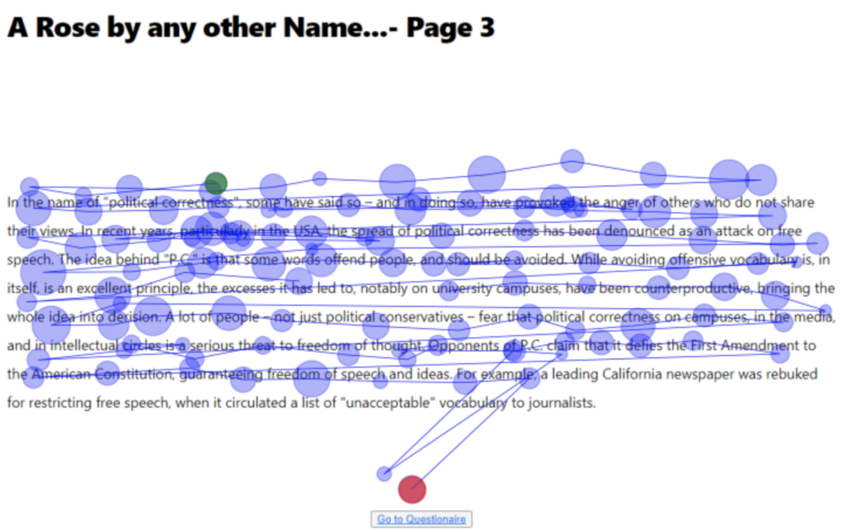
In order to simulate a realistic e-learning environment, we created a web-based document reading platform utilizing technologies like Next.js, React.js, and Express.js. This online platform facilitated the development of an interactive and dynamic learning environment, providing participants with a smooth user experience. During the experiment, we employed a Tobii 4C eye tracker with a professional license, affixed to a display monitor, to monitor the participants' eye movements as they read the documents. Moreover, the Empatica E4 wristband was used to capture their physiological responses throughout the reading activity. The experiment was organized into two distinct sessions, allowing participants to take a 10-minute break after reviewing the first seven documents. This break was provided to help reduce potential fatigue and ensure that participants remained attentive and engaged for the entire duration of the experiment [160].

3.4.2 Methodology

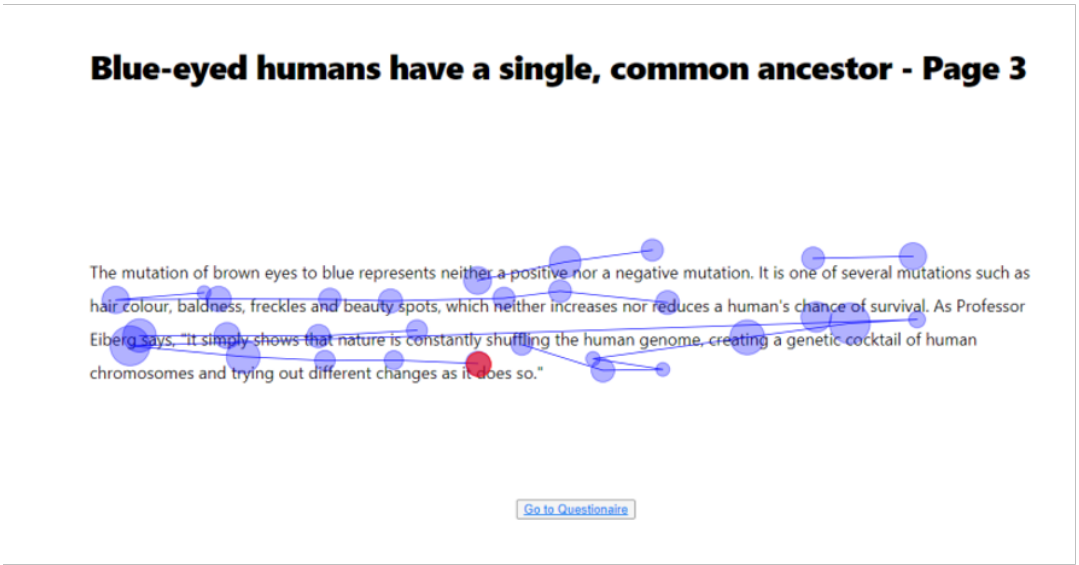
This system employed a comprehensive methodology incorporating gaze tracking and physiological signal recording during reading sessions to assess readers' affective states, including engagement, arousal, and valence. The Tobii 4C remote eye-tracker with a pro license captured gaze coordinates and pupil diameters at a sampling rate of 90 Hz. Physiological data, including 3-axis acceleration (ACC) at 32 Hz, blood volume pulse (BVP) at 64 Hz, electrodermal activity (EDA) at 4 Hz, skin temperature (TEMP) at 4 Hz, and heart rate (HR) at 1 Hz, were acquired using the E4 wristband sensor. Engagement, valence, and arousal were detected using gaze data, E4 data, and a combination of both. The section methodology followed by the subsections are based on our published research [138, 160].

Data pre-processing

The gaze data included raw eye data such as the coordinates for both the left and right eyes and pupil diameter, all recorded at a frequency of 90 Hz. To prepare this raw data for further analysis, preprocessing steps were implemented to eliminate outliers and reduce noise, followed by a sliding window technique that divided the data into 30-second segments with a 50% overlap. Figure 28 illustrates the fixations and saccades of a participant who evaluated various documents as extremely engaging (Figure 28a) and least engaging (Figure 28b). Similarly, physiological data such as EDA, BVP, TEMP, HR, and 3-axis ACC were collected using the Empatica E4 device, with sampling frequencies of 4 Hz, 64 Hz, 4 Hz, 1 Hz, and 32 Hz, respectively. The gaze and physiological data were synchronized through their timestamps, and preprocessing was also conducted for the physiological signals, which included noise reduction and applying a sliding window technique with a 30-second duration and a 50% overlap for data segmentation for analysis [160].



(A) Highly Engaging Document



(B) Less Engaging Document

FIGURE 28: Scanpaths of gaze data for documents rated as highly engaging (left) vs. less engaging (right) by a participant.

Model Architectures

This study employed five deep learning architectures to detect engagement, arousal, and valence: Transformers, Fully Convolutional Networks (FCNs), an Encoder network, Residual Networks (ResNets), and Inception networks. These architectures were chosen for their ability to automatically learn relevant features from raw multimodal data (gaze and physiological signals), reducing the need for manual feature engineering [160].

Transformer Network: Adapted from natural language processing, the Transformer network uses self-attention mechanisms to capture long-range temporal dependencies within sequential gaze and physiological data. Input consists of individual and concatenated gaze and physiological data, with positional encoding to maintain temporal order. Self-attention layers process the data, followed by fully connected feed-forward networks for feature extraction and classification. This architecture excels at identifying subtle temporal patterns in multimodal data.

Fully Convolutional Network (FCN): The FCN uses three convolutional blocks per signal (gaze and physiological), followed by global average pooling and fully connected layers for classification. Convolutional layers extract features, and the fully connected layers perform the final classification.

Encoder Network: Similar to the FCN, the encoder network includes an attention layer between the convolutional and fully connected layers. This attention mechanism allows the network to focus on the most relevant features within the input sequence, improving feature extraction and classification accuracy.

Residual Network (ResNet): The ResNet architecture employs residual blocks with shortcut connections to mitigate the vanishing gradient problem during training. These residual blocks effectively capture complex temporal patterns and dependencies in the time series data.

Inception Network: The Inception network builds upon the ResNet architecture by incorporating inception blocks in addition to residual blocks. Inception blocks use parallel convolutional layers with varying kernel sizes to capture multi-scale patterns, enhancing the network's ability to handle complex data.

Classification

The deep neural networks were developed to predict the engagement, arousal, and valence reactions as reported by the participants. For engagement, participants were instructed to rate their experience on a scale of '1' to '5'. These scores were subsequently divided into two separate categories: *Low* and *High*. Scores ranging from '1' to ‘3’ were classified as *Low* engagement, whereas scores of ‘4’ and ‘5’ were viewed as *High* engagement. A similar methodology was applied for predicting the arousal and valence reactions. The arousal and valence responses provided by participants were segmented into four quadrants, each linked to a specific emotional state. The quadrants are as follows [160]:

- 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.

Evaluation Protocols

Model performance was evaluated using user-independent (leave-one-participant-out, or LOPO) and user-dependent approaches. LOPO iteratively trained on all but one participant’s data, testing on the excluded participant, and averaging the results. The user-dependent approach similarly used iterative training and testing, but on individual participants’ data, averaging results across participants to account for individual differences [160].

3.4.3 Results

Deep learning models predicted engagement, valence, and arousal using two training methods: user-independent and user-dependent. These predictions were based on a binary classification task, with ground truth derived from participant survey responses. To provide a more comprehensive emotional analysis, a four-class classification structure was also implemented, segmenting valence and arousal responses into four emotional quadrants. The results along with the subsequent sections are based on our published research [138, 160].

Model Performance: User-Independent Approach

Model performance, evaluated using accuracy and F1-score, is presented in Table 3.13 for all models trained using a user-independent approach. Evaluation metrics were calculated for individual modalities (eye-tracking and wristband data) and their combination to assess the benefit of multimodal data fusion. The Transformer model consistently outperformed other models in terms of accuracy and F1-score for engagement, valence, and arousal prediction. For engagement, the combination of eye-tracking and wristband data yielded the highest accuracy (80.38%) and F1-score (78.73%). Similarly, the Transformer model achieved superior performance for valence (71.28% accuracy) and arousal (73.98% accuracy) prediction compared to all other models [160].

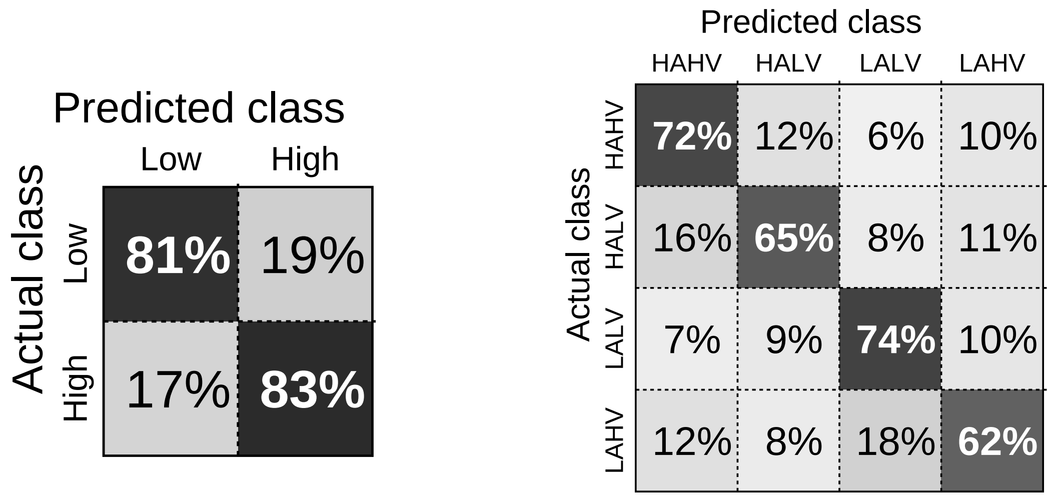
TABLE 3.13: Summary of evaluation metrics using **user-independent** approach (LOPO)

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**

ResNet and Inception models also demonstrated reasonably good performance. Importantly, multimodal data fusion consistently improved prediction accuracy across all three target variables, highlighting the benefits of a multimodal approach. However, using wristband data alone resulted in lower performance compared to eye-tracking data or the combined modality. This difference may be attributed to the lower sampling frequency of the Empatica E4 wristband, potentially leading to a less detailed dataset and impacting the accuracy of emotion and engagement prediction. Figure 29 presents confusion matrices visualizing the performance of the binary engagement classification (Figure 29a) and the four-class emotion classification (Figure 29b) models using the user-independent approach [160].

Model Performance: User-Dependent Approach

The research also emphasizes that both engagement and emotions are highly personal experiences that can vary significantly among different people, even when



(A) Engagement Classification (Binary)

(B) Emotion Classification (Four-Class)

FIGURE 29: Confusion matrices for user-independent classification. (a) shows binary engagement classification (low/high). (b) shows four-class emotion classification (HAHV: High Arousal High Valence, HALV: High Arousal Low Valence, LALV: Low Arousal Low Valence, LAHV: Low Arousal High Valence).

they are engaged in the same activities. Each person has their own distinct traits and patterns of emotional and cognitive reactions, which contribute to a unique profile of their engagement and emotional responses. As a result, using a generalized model to detect and predict these experiences may not be effective due to its lack of personalization. To tackle this issue and customize our predictive models to reflect individual experiences, we have adopted a user-dependent approach. This strategy prioritizes personalization, integrating specific user traits and patterns into the predictive models to design a distinct model for each person. This differs from the user-independent approach, which generally builds models based on pooled data from various users without recognizing individual variations [160].

Table 3.14 shows that the user-dependent approach consistently improved prediction accuracy for engagement, valence, and arousal across all models compared to the user-independent approach. The ResNet model, using combined modalities, achieved the highest accuracy: 93.56% for engagement, 90.62% for arousal, and 88.70% for valence. These results highlight the importance of personalization in improving the accuracy of emotion and engagement detection. Figures 30a and 30b present confusion matrices for the binary engagement and four-class emotion classification models, respectively, using the user-dependent approach [160].

3.4.4 Discussion

This section explores the primary contributions, results, and insights obtained from this study. It includes both the theoretical significance of the research and the potential practical applications derived from it. In addition, this section reflects on the innovative analysis method employed and the importance of the research question

TABLE 3.14: 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

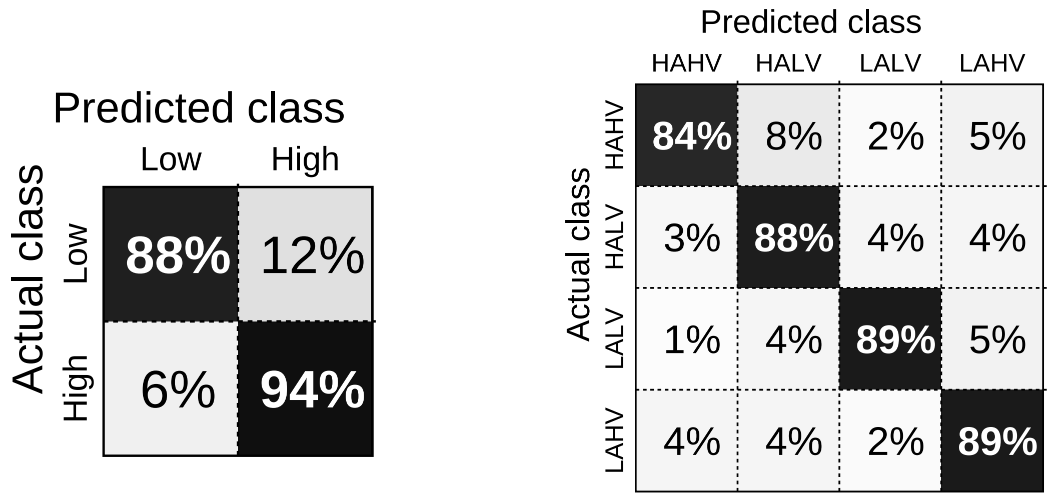
in relation to current discussions in the field. The discussion section builds upon our published findings [138, 160].

Experimental protocol for engagement and emotion detection

This study uniquely integrates engagement and emotion (arousal and valence) detection during a reading task, a more complex scenario than video-based emotion recognition. We used unobtrusive data collection (eye-tracking and physiological signals from an Empatica E4 wristband) to capture both engagement and emotional responses. Careful document selection via sentiment analysis ensured diverse emotional responses (happiness, anger, boredom, calmness) across various genres, while a controlled presentation format (one paragraph at a time) minimized data noise, enhancing data reliability and the validity of our findings [160].

User-Independent and User-Dependent prediction using deep learning models

The preprocessed data from eye-tracking and wristband were used to train state-of-the-art time series classification models for engagement, valence, and arousal



(A) Engagement Classification (Binary)

(B) Emotion Classification (Four-Class)

FIGURE 30: Confusion matrices for user-dependent classification. (a) shows binary engagement classification (low/high). (b) shows four-class emotion classification (HAHV: High Arousal High Valence, HALV: High Arousal Low Valence, LALV: Low Arousal Low Valence, LAHV: Low Arousal High Valence).

detection. While the models were quite successful at identifying engagement levels, predicting valence and arousal proved more challenging. Combining data from both devices significantly improved accuracy, an outcome that highlights the value of multimodal data collection. Two approaches were used: a user-independent approach (Leave-One-Participant-Out cross-validation) and a user-dependent approach, allowing comparison between generalized and personalized models. Under the user-independent approach, the Transformer network excelled, particularly when combining features, though valence and arousal predictions required further refinement. Interestingly, gaze features proved stronger predictors than physiological signals when modalities were used individually, potentially due to the wristband's lower sampling frequency [160].

The user-dependent approach significantly improved performance across all models, with substantial accuracy gains (13% for engagement, 19% for arousal, and 15% for valence). The ResNet model achieved the best overall performance. Among the models we tested, ResNet consistently performed best. The success of the personalized approach underscores the importance of considering individual differences when designing systems that aim to understand and respond to user experiences. This has significant implications for improving user engagement in various applications, from e-learning platforms to marketing campaigns [160].

Engagement and Emotion Correlation

This study investigated the correlation between engagement and emotions (arousal and valence) during a reading task, using 1-5 scales for self-reported engagement, valence, and arousal. Figure 31a shows a correlation matrix heatmap and Figure 31b shows a boxplot visualizing the distribution of engagement across various emotional categories. The heatmap reveals strong positive correlations between engagement and arousal (0.68) and moderate positive correlations between engagement and valence (0.44). The boxplot shows that high-arousal emotions (positive or negative) correlate with higher engagement, while low-arousal emotions correlate with lower engagement [160].

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 which could be crucial for educators and instructional designers to develop effective strategies [160].

(A) Correlation matrix heatmap showing the linear relationship between engagement, arousal, and valence.

(B) Boxplot visualizing the distribution of engagement across various emotional categories.

FIGURE 31: Correlation matrix heatmap and boxplot for engagement and emotional categories.

3.4.5 Limitations and Future Work

Although the discussed study makes significant contributions to user engagement and emotional reactions during reading activities, several limitations and challenges exist that could be addressed in future research. These include:

Generalization Across User Groups: The research proposes a user-independent generalized framework; however, it may still be difficult to ensure that the results apply universally to a variety of user groups. Users differ greatly in terms of preferences, demographics, and personal characteristics, which may affect the models' effectiveness across distinct user segments. Future efforts could involve gathering data from a more varied population to enhance the models' generalizability and customize them for particular user groups [160].

Subjectivity in Emotion Assessment: Evaluating emotions is often inherently subjective. Individuals may have different interpretations and expressions of emotions, and depending solely on self-reported feelings or external cues like emojis may not fully capture the complexities of users' emotional experiences. Ensuring the reliability and validity of emotional evaluations remains a persistent challenge. Future investigations could examine the inclusion of multimodal data, such as physiological signals and facial expressions, to supplement self-reported feelings and enhance assessment accuracy [160].

Model Robustness and Generalization: While the study underscores a comprehensive assessment of deep learning models, their robustness in real-world situations and ability to generalize across various reading contexts may pose a challenge. The models' performance observed in controlled experimental environments might not entirely represent their effectiveness in diverse and natural reading settings. Future research could prioritize testing the models in actual environments and incorporating adaptive features to accommodate unpredictable user behaviors [160].

Correlation vs. Causation: The exploration of the link between user engagement and emotions is important, yet establishing a causal connection is complex. Though the study might indicate relationships, it may not conclusively determine whether shifts in engagement lead to changes in emotions or the opposite. Future investigations could implement experimental methodologies, such as controlled interventions, to more effectively confirm causal connections between engagement and emotional states [160].

Long-term User Engagement: By concentrating on short-term reading tasks, the study may limit its understanding of long-term user engagement and emotional trends. Gaining insights into how these dynamics develop over extended durations could offer a more comprehensive perspective on user experiences. Future research could focus on longitudinal studies to evaluate how engagement and emotional reactions evolve over time and in various contexts [160].

Tackling these limitations and challenges could enhance the study's influence and contribute meaningfully to the ongoing discussion surrounding user engagement, emotions, and the use of deep learning in understanding human behavior during reading processes.

3.4.6 Conclusion

This study enhances the understanding and detection of engagement and emotions while reading, employing a multimodal strategy that integrates gaze tracking and physiological responses from participants. A variety of deep learning models were applied to analyze and process the collected multimodal data, enabling a thorough comparison of the effectiveness of these models in handling such complex tasks. An experimental framework was established that utilizes a non-intrusive multimodal method to capture user responses related to engagement, valence, and arousal. We applied both a generalizable user-independent approach and a personalized user-dependent approach to create tailored predictions. This comparative research shed light on the performance of these models in complex tasks. The investigation explored the relationship between user engagement and the emotions elicited during reading activities. This knowledge offered valuable insights into how emotions and cognitive engagement are interrelated during reading tasks, contributing to the development of strategies aimed at improving user engagement [160].

3.5 Readability Analysis and Detection

This section is adapted from our published work *"Generating Heatmap for Unknown Documents towardsReadabilityMeasurement"* [158]. Eye tracking is one of the most promising ways to analyze the visual patterns of reading and it could be done in real time, providing details about the participant's reading behavior [78]. By assessing the reading behavior it is possible to estimate the readability of the document [25]. The content features also play a crucial role in measuring the readability of the document. The complexity, familiarity, and length of the words are some features that could determine the difficulty of the document a participant is reading. The difference in reading patterns could be very well observed among native language speakers and learners. Learners tend to spend more time reading a certain document to completely acquire the information. On the other hand, for a native speaker, the words or sentence constructions will be more familiar and the variance in reading pattern could be observed. So the main motivation behind this research was to measure the readability of a document in a more personalized manner rather than just considering the linguistic complexity, syntactic structure, and surface text properties. The incorporation of eye gaze features of different users along with the text features would provide a much better picture of user perception of the text.

The pattern of reading for a document that is difficult to understand and easy to understand could vary a lot. Based on the interest in a document, the attention and involvement of a participant reading a document would change. For content which comprises of long and unfamiliar words as well as complex sentence construction, the user tends to re-fixate the area again which could increase the fixation duration. As a matter of fact, this would vary depending on the user's language skill too. The text-related features alone could not include the cognitive load associated with the reading behavior.

The integration of text features along with gaze features could improve the results of readability prediction as it also covers the cognitive processing in the background and determines the grade of the document [179]. The text difficulty could be assessed in this way and could be further utilized to simplify the text, which would make it readable to a wider audience. The digitized versions of text are becoming more popular and the use of printed books and papers is declining over time. To enhance the traditional methods or formulas for measuring readability, new features that considers the user behavior also should be included, which would result in better estimation of reading difficulty.

The integration of natural language processing (NLP) tools and machine learning models could be used for readability assessment, which provides information about the difficulties in reading [13]. Multiple regressions and more fixations over a region in a document could mean that it consists of either complex words or complicated sentence construction, which makes it difficult for the reader to understand the context.

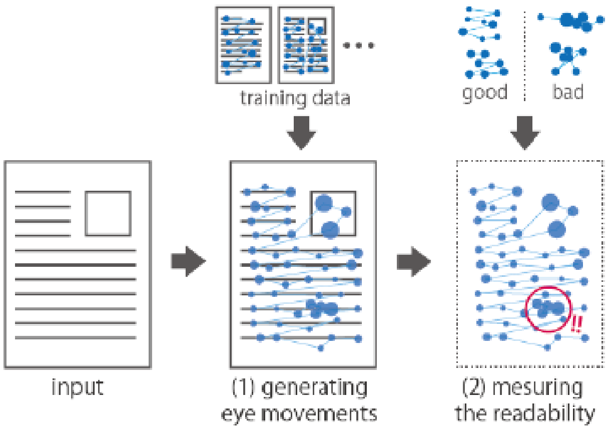


FIGURE 32: The concept image of generating eye movements and measuring readability.

The main objective of this research is to predict readability ratings for documents based on their level of difficulty [158]. To achieve this, two studies were conducted: the first focused on generating artificial eye movements (Figure 32) and fixation heatmaps for unknown documents, while the second involved integrating real eye-tracking data with textual features for readability prediction. In this context, the following research questions are addressed:

RQ1. How can artificial eye movements be generated for unknown documents, and how can fixation heatmaps be used to estimate readability?

RQ2. How can fixation heat values be calculated for individual words, and how can these values be used as labels for readability prediction?

RQ3. How can machine learning algorithms be utilized to predict readability ratings for documents based on their difficulty levels?

RQ4. How can eye movement patterns, both artificial and real, be visualized for documents with varying readability ratings, and how can the relationship between reading behavior and document content be estimated?

3.5.1 Data Collection

Two studies were conducted using the Tobii 4C to record participants' eye movements while reading documents. The eye tracker was mounted on a desktop monitor, and participants calibrated the device before reading each document to ensure accurate data collection.

Study 1: Pilot Experiment

The first study involved 20 college students reading six documents arranged in ascending order of difficulty. The first document served as a sample to familiarize participants with the reading task and the associated questionnaires. Each document was followed by three multiple-choice questions directly related to its content, designed to ensure participants' engagement and comprehension while reading. The inclusion of these questions minimized skimming behavior and provided insights into participants' understanding capabilities. Eye movement data was recorded as participants read the documents and answered the questions.

Study 2: Extended Experiment

The second study involved 10 master's students analyzing 15 documents chosen from BBC news articles covering various topics such as business, technology, politics, sports, and entertainment. This category-based selection ensured a broad range of reading difficulties. To enhance precision, the documents were presented one paragraph at a time, with four surrounding dots acting as visual boundaries to restrict fixations within the reading area. Participants were asked to focus on these dots prior to reading to improve tracking accuracy [158].

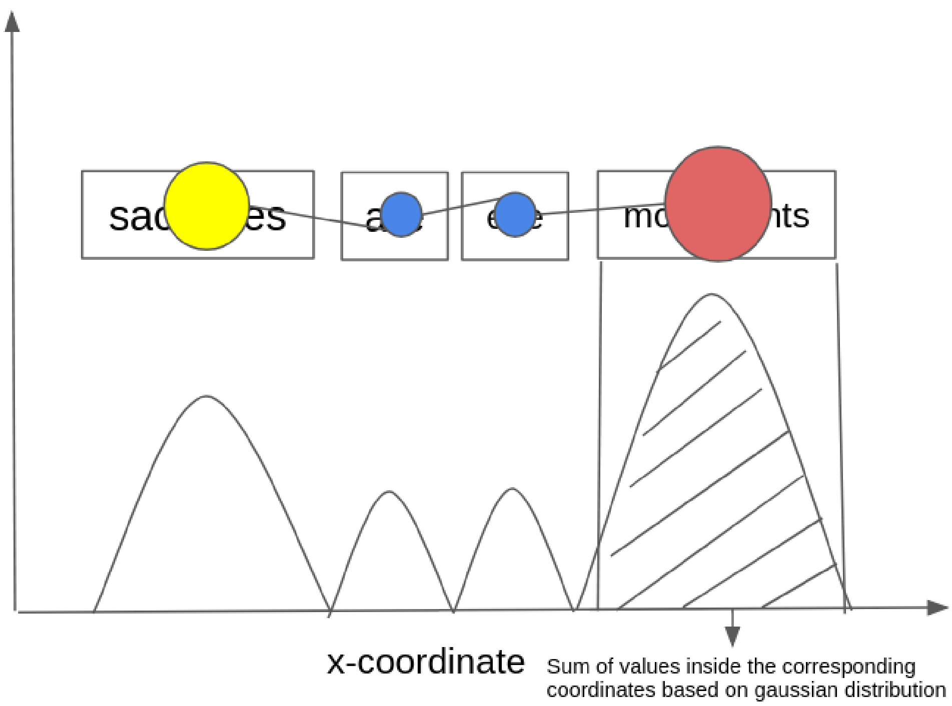


FIGURE 33: The fixation heatmap values corresponding to each word coordinate is summed to find the fixation heat value.

3.5.2 Methodology

The methodology for predicting readability involved collecting eye movement data, extracting features from document content, and applying machine learning models to predict fixation durations per word and classify document difficulty. Both studies followed a similar pipeline, with variations in the features used, regression models applied, and additional tasks such as difficulty classification in the second study.

Ground Truth Recording

In both studies, raw gaze data recorded by the Tobii Eye Tracker 4C was processed to detect fixations and saccades. Heatmaps were generated to represent the distribution of fixations over the document. Fixation durations per word were calculated by summing the heatmap values within the rectangular bounding box of each word, defined by its top-left and bottom-right coordinates. This summed value, referred to as the *fixation heat value*, was used as the ground truth for prediction as shown in Figure 33 [158]. To account for variations across participants, the mean fixation heat value for each word was computed by aggregating data from all participants.

Feature Calculation

Features were extracted from the document content to predict fixation durations and classify document difficulty. Across both studies, the *length of each word* was used as a primary feature, based on the assumption that longer words tend to have higher fixation durations. Additional features varied between the studies:

- **Study 1:**
  - Word embeddings were extracted using Google's pre-trained Word2Vec model, resulting in a 300-dimensional feature set. Principal Component Analysis (PCA) was applied to reduce the dimensionality, and the top five significant features were selected.
  - Stop words were removed, and all words were converted to lowercase before feature extraction.
- **Study 2:**
  - Lexical and syntactical features, such as *log-likelihood properties* (e.g., lexical, syntactical, and total surprisal) and *ambiguity*, were computed using an incremental probabilistic top-down parser.
  - *Parts of speech (POS) tags* (e.g., NN, VB, IN, JJ) were extracted for each word and included as features.
  - Gaze features, such as fixation duration, saccade length, regression length and regression ratio, were also utilized to capture reading behavior [158].

Machine Learning Models

Two different machine learning tasks were performed across the studies:

1. **Fixation Duration Prediction:**
   - **Study 1:** Support Vector Regression (SVR) with an RBF kernel was applied to predict fixation durations per word based on the extracted features.

- **Study 2:** A Random Forest Regression model was used, leveraging the diverse set of lexical, syntactical, and gaze features for prediction.

2. **Document Difficulty Classification (Study 2):** A Support Vector Machine (SVM) was used to classify documents into four difficulty levels based on participant ratings:

   - **1:** Very hard to read
   - **2:** Moderately hard to read
   - **3:** Easy to read
   - **4:** Very easy to read

   The SVM model was trained using text features, gaze features, and a combination of both feature sets to evaluate the impact of each on classification performance.

3.5.3 Results and Analysis

The results of both studies are presented below, focusing on the prediction of fixation heat values, the generated heatmaps, and the analysis of document readability.

Study 1: Fixation Heat Value Prediction

A Support Vector Regression (SVR) model and a Linear Regression model were used to predict fixation heat values per word. The models were evaluated using Leave-One-Participant-Out (LOPO) and Leave-One-Document-Out (LODO) approaches. The results, shown in Table 3.15, indicate that word length as a single feature contributed significantly to the model's performance. The inclusion of word embeddings and traditional text features (e.g., word frequency and POS tags) slightly improved the results, with the linear regression model using LOPO and traditional features achieving the highest goodness score of 0.503.

TABLE 3.15: Goodness scores using different features and methods.

Features	LOPO (SVR)	LODO (SVR)	LOPO (LR)	LODO (LR)
Word Length	0.469	0.451	0.4838	0.4735
Word Length + Embeddings	0.478	0.454	0.4865	0.4788
Traditional Text Features	0.4912	0.471	0.503	0.4934

Figure 34 illustrates the generated scan path over a document. The scan path colors represent fixation durations, with blue indicating values below 200 ms and red indicating values above 600 ms. The boxplot in Figure 35 demonstrates the relationship between word length and fixation heat values, confirming that longer words tend to have higher fixation durations. Despite these findings, the calibration errors of the eye tracker and the limited feature set impacted the overall accuracy of the predictions.

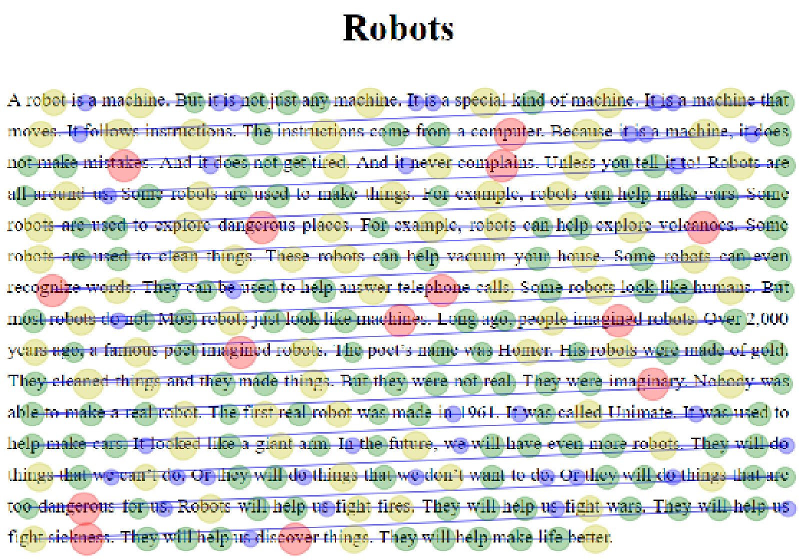


FIGURE 34: Generated scan path over the document based on predictions.

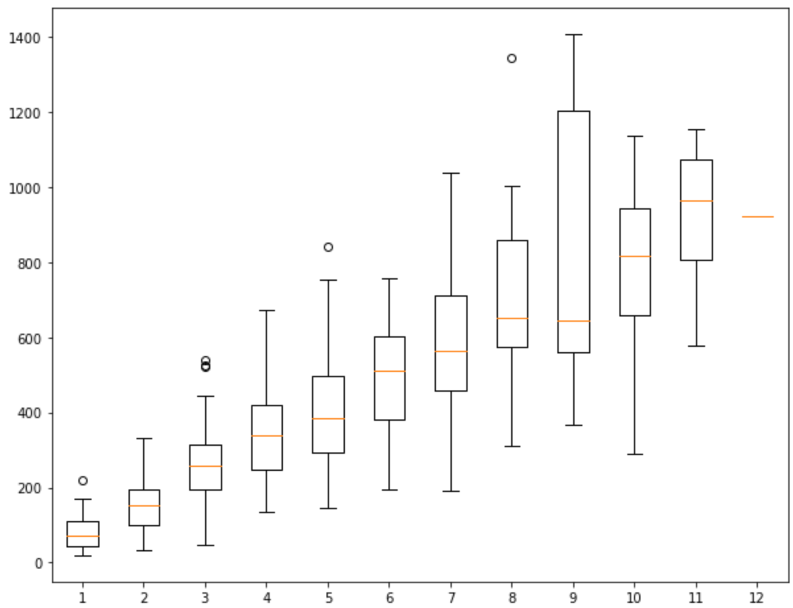


FIGURE 35: A boxplot showing the fixation heat values with respect to the word lengths(x-axis : word length; y-axis : fixation heat values)

The generated heatmaps (Figure 36) showed some alignment with the ground truth, particularly in regions with higher fixation durations. However, the predictions were limited by the small number of text features used, and the addition of more lexical and syntactical features could improve accuracy.

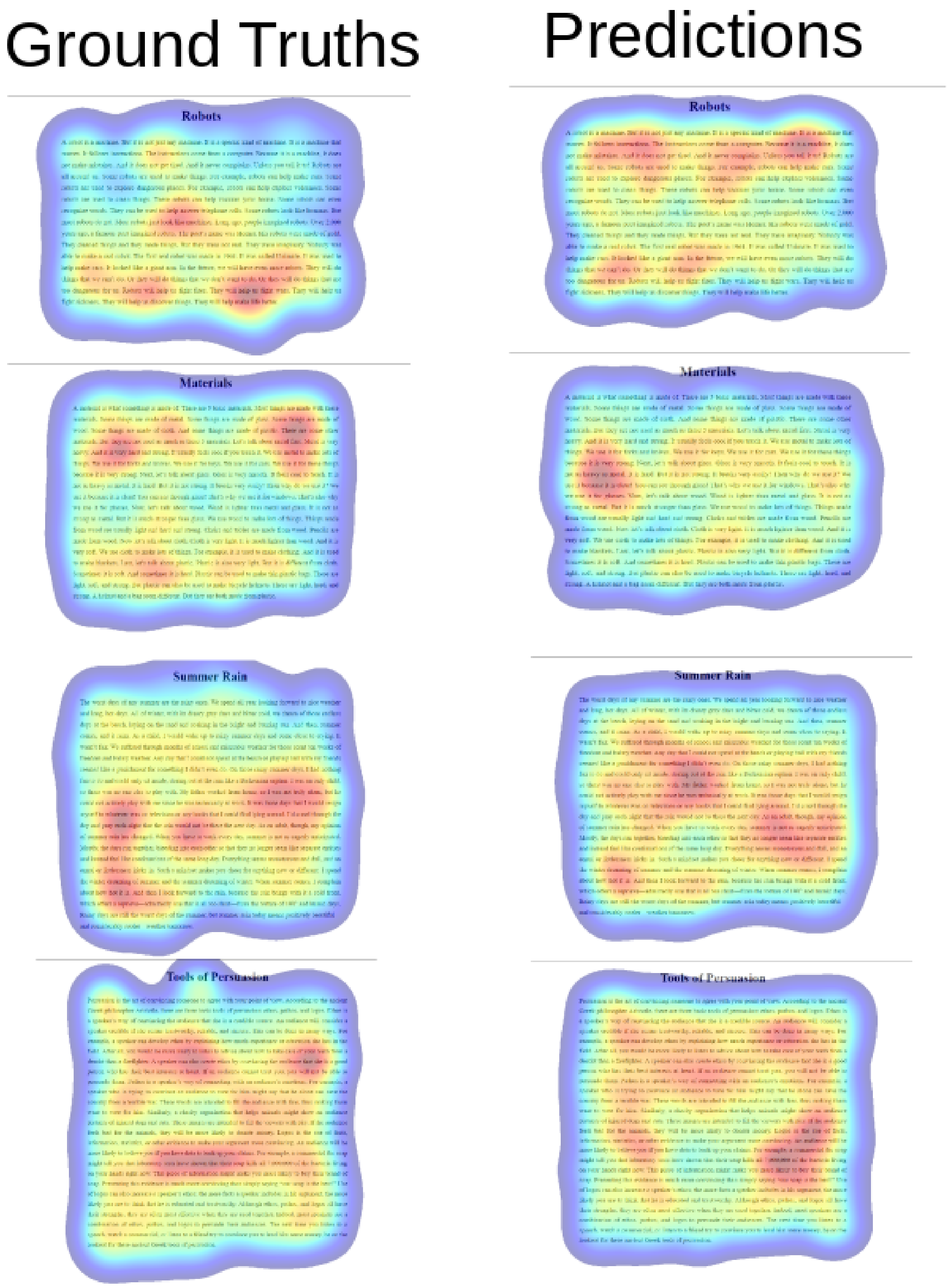


FIGURE 36: The ground truth heatmap of the documents along with the predictions obtained.

Study 2: Fixation Heat Value Prediction and Heatmap Analysis

In the second study, a random forest regression model was trained using the Leave-One-Document-Out (LODO) approach. A grid-search cross-validation technique was utilized to fine-tune hyperparameters. The model achieved a mean regression score (R^2) of 0.757 and a normalized root mean squared error (NRMSE) of 0.282. An analysis of feature importance indicated that word length was the most critical feature, boasting an importance value of 0.972. Using the predicted fixation heat values, heatmaps were generated for each document to identify the main focus areas. Figure 37 shows the ground truth and predicted heatmaps for three documents. The model successfully captured the primary focus areas (red regions) in the documents, aligning with the ground truth to a reasonable extent. The addition of lexical and syntactical features, such as log-likelihood properties and POS tags, significantly improved the accuracy of the generated heatmaps compared to Study 1. This improvement highlights the importance of incorporating diverse text features for better readability prediction.However, variations in predictions were observed due to differences in participant behavior and the complexity of the documents [158].

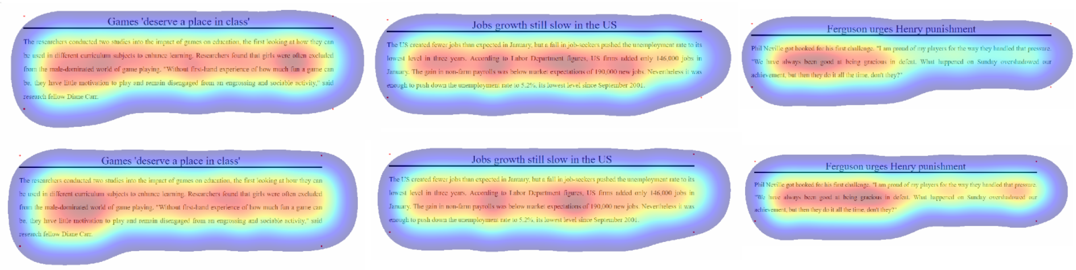


FIGURE 37: Three examples of heatmap generation (top: ground truth, bottom: generated)

In addition to fixation heat value prediction, a Support Vector Machine (SVM) was used to classify document difficulty based on participant ratings. The ratings were divided into four classes: *1: very hard to read*, *2: moderately hard to read*, *3: easy to read*, and *4: very easy to read*. The model was trained using text features, gaze features, and a combination of both. To optimize the SVM hyperparameters, a *GridSearch CV* was performed, resulting in the best parameters: C: 1, *gamma*: 0.5, and *kernel*: *rbf*. Class weights were balanced to account for the uneven distribution of ratings, as ratings “2” and “3” were more frequent than “1” and “4”.

The results of the SVM classifier for three cross-validation methods are shown in Table 38a. The Leave-One-Document-Out (LODO) approach, which is most suitable for predicting the readability of unknown documents, achieved a mean accuracy of 51% using text features and 52% with the combination of text and gaze features. However, due to the limited data for rating “1,” predictions were primarily in the range of “2-4.”

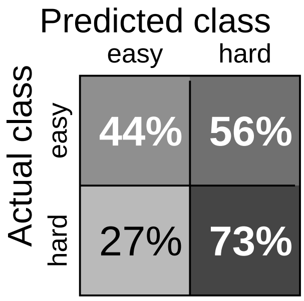
To further simplify the task, a binary classification was performed, grouping ratings “1” and “2” as *hard to read* and “3” and “4” as *easy to read*. The binary SVM classifier achieved a mean accuracy of 63%, as shown in the confusion matrix in Figure 38b. This demonstrates the potential of combining text and gaze features for readability classification.

3.5.4 Discussion

The two studies aimed to predict document readability by leveraging fixation heatmaps and gaze features. Study 1 served as a pilot study, focusing on generating fixation

Features	**LOPO**	**LODO**
Text Features	0.39	0.51
Gaze Features	0.42	0.48
Combined	**0.50**	**0.52**

(A) Mean accuracy.



(B) Confusion matrix.

FIGURE 38: (a) Mean accuracy for different cross-validation methods using the *SVM* classifier (4-class), and (b) Confusion matrix for binary *SVM* classifier with combined features.

heatmaps for unknown documents using word length as the primary feature. However, several challenges emerged, including calibration errors that caused vertical misalignment of fixation coordinates with the content lines. This misalignment hindered the accurate extraction of fixation durations per word. To address this, fixation heatmaps were generated for each document, and the sum of heatmap values within word boundaries was used to compute fixation heat values. Despite these efforts, the reliance on word length alone and the use of Support Vector Regression (SVR) yielded suboptimal results, as the fixation patterns varied significantly across participants.

Study 2 built upon the limitations of Study 1 by introducing several methodological improvements. First, the number of participants was reduced, and the number of documents per participant was increased to minimize participant-specific biases. Additionally, documents were categorized into specific topics, enabling a more structured analysis. The experimental setup was refined by displaying documents paragraph by paragraph, increasing font size and line spacing, and introducing boundary markers to improve gaze data accuracy. Although participants often forgot to fixate on the boundary markers, the overall data quality improved significantly compared to Study 1.

A key advancement in Study 2 was the implementation of a line-break detection algorithm to realign fixation coordinates with the content lines, effectively addressing the vertical misalignment observed in Study 1. This algorithm successfully aligned fixations for most participants, as shown in Figure 39, resulting in more reliable fixation heatmaps.

In terms of feature selection, Study 2 expanded beyond word length by incorporating lexical and syntactical features, such as log-likelihood probabilities (e.g., surprisal and ambiguity). These features demonstrated a stronger correlation with participant ratings, improving the model's performance. Unlike Study 1, where word embeddings failed to enhance predictions, the inclusion of diverse text features in Study 2 significantly improved the accuracy of fixation heat value predictions and heatmap generation.

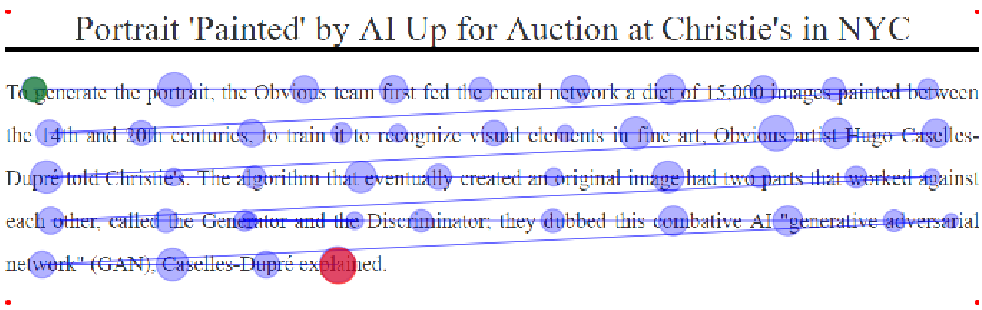


FIGURE 39: The realigned fixations and saccades based on the line break algorithm.

For document difficulty classification, Study 2 employed an SVM classifier, which outperformed the regression models used in Study 1. However, the imbalanced distribution of participant ratings (e.g., very few ratings of '1: very hard to read') remained a challenge, leading to biased predictions favoring the '2-3' range. While combining text and gaze features did not yield the expected improvement, gaze features such as mean fixation duration and saccade length showed positive correlations with difficulty ratings for some participants.

Overall, Study 2 addressed the key limitations of Study 1 by improving data collection accuracy, refining experimental design, and incorporating more diverse features. These advancements resulted in more reliable fixation heatmaps and better readability predictions, highlighting the importance of addressing calibration errors, participant biases, and feature diversity in eye-tracking studies.

3.5.5 Limitations and Future Work

Despite the improvements in Study 2, several limitations remain. First, the imbalanced distribution of participant ratings, particularly the scarcity of 'very hard to read' ratings, introduced bias in the difficulty classification task. This imbalance limited the model's ability to accurately predict documents at the extremes of the difficulty scale. Expanding the rating scale or increasing the diversity of document difficulty levels could address this issue in future studies.

Second, while the line-break detection algorithm successfully realigned fixation coordinates for most participants, a few recordings still exhibited misalignments due to calibration errors. These residual inaccuracies may have affected the fixation heat value predictions and subsequent heatmap generation. Future research could explore more robust calibration techniques or adaptive algorithms to minimize misalignments and improve the accuracy of fixation data.

Third, the combination of text and gaze features did not yield the expected improvements in classification performance. This suggests that the interaction between these features may require more sophisticated modeling techniques or additional preprocessing to fully capture their relationship with document readability. Future studies could investigate advanced modeling approaches, such as deep learning architectures or feature fusion techniques, to better integrate text and gaze features for improved performance.

Finally, participant behavior, such as forgetting to fixate on boundary markers or re-reading document titles, introduced variability in the gaze data. Although the experimental design in Study 2 reduced these issues compared to Study 1, further refinements in task instructions and experimental setup are necessary to ensure consistent data quality. Future work could focus on designing more structured experimental protocols and providing clearer task instructions to minimize variability in participant behavior.

These limitations highlight areas for future work, including addressing data imbalances, improving calibration techniques, and exploring advanced modeling approaches to better integrate text and gaze features.

3.5.6 Conclusion

This research focused on assessing document readability using fixation heatmaps and gaze features, with the goal of generating artificial eye gaze for unknown documents. The pilot study revealed significant challenges, including calibration errors and limited feature diversity, which impacted the accuracy of fixation heat value predictions. Study 2 addressed these issues by refining the experimental design, improving data collection accuracy, and incorporating additional features such as surprisal and ambiguity. These enhancements resulted in more reliable fixation heatmaps and better insights into document readability.

In Study 2, difficulty rating prediction was introduced as a classification task using an SVM model. The model achieved a mean accuracy of 51% using text features and a combination of text and gaze features. However, the imbalanced distribution of ratings, with very few 'very hard to read' ratings, limited the model's ability to predict extreme difficulty levels. Binary classification (easy vs. hard) improved accuracy to 63%, but the results highlighted the need for more balanced data and additional features to improve performance. Features such as saccade length, regression ratio, and fixation duration showed positive correlations with difficulty ratings, providing valuable insights into participant reading behavior. Despite these advancements, the integration of text and gaze features did not yield the expected improvements in classification performance, primarily due to participant-specific biases and variability in gaze behavior. The line-break detection algorithm successfully realigned fixations, improving the accuracy of gaze feature extraction. However, challenges such as limited data for extreme difficulty ratings persisted.

Chapter 4

Personalized Interventions

This chapter explores various personalized intervention strategies designed to enhance cognitive performance, manage distractions, and improve user engagement in diverse contexts. Personalized interventions are essential for tailoring solutions to individual needs, enabling systems to adapt dynamically to users' cognitive and emotional states [182]. The goal of this research is to develop and evaluate methods that leverage controlled distraction, cognitive training, emotional engagement, and real-time adaptive systems to address challenges in attention, learning, and user interaction.

The chapter is organized into several sections, each focusing on a specific type of intervention. The first section introduces distraction management in reading, where controlled distraction techniques are used to study and mitigate the effects of distractions on reading behavior [187]. This is followed by a section on cognitive control training for digital distraction management, which explores methods to help users improve their focus and reduce the impact of digital distractions [159].

The next section focuses on interventions targeting emotion and engagement, where models and interactive interfaces are developed to enhance user engagement and emotional connection during tasks [160]. This is followed by a section on a real-time gaze-based adaptive learning system, which leverages real-time eye-tracking data to create personalized learning experiences by dynamically adapting content based on user behavior and cognitive load [156]. The section concludes with an analysis of the impact of these interventions, including their implications for cognitive load, visual effort, and overall user experience.

The research presented in this chapter provides a foundation for designing intelligent systems that can adapt to individual needs, offering significant potential for applications in education, productivity, and human-computer interaction. The following sections of this chapter are based on collaborative research with students, each contributing to specific aspects of the work. Section 4.1 presents research conducted with Pramod Vadiraja, who was actively involved in data collection, analysis, and scientific writing for publication. My contributions to this section included co-supervision, technical analysis, and preparation of the manuscript for publication. Section 4.2 is based on work carried out with Shrishti Jagtap, who assisted with data collection and participant recruitment. For this study, I contributed to the data analysis, modeling, evaluation, and preparation of the scientific article. Section 4.3 includes research developed in collaboration with Akshay Palimar Pai, who provided assistance in application development. Across all studies, my role involved conceptualization, technical analysis and guidance, and publishing the resulting scientific articles..

4.1 Controlled Distraction Intervention in Reading

This section is adapted from our published work *"Effects of Counting Seconds in the Mind while Reading"* [187]. Distraction and attention are two opposing cognitive processes that significantly impact task performance, particularly in reading [56]. While tasks with high perceptual load can reduce distraction by fully engaging cognitive resources, tasks with high cognitive load on working memory are more susceptible to attention deterioration [189]. Distractions during reading can arise from various factors, such as mind wandering, information overload, or lack of interest. Media multitasking, for instance, has been linked to poor cognitive task performance, potentially due to increased sensitivity to distractions [124].

FIGURE 1: Scan-paths of gaze data during reading without distraction (top) and with counting distraction (bottom).

In this study, a novel controlled distraction intervention mechanism is introduced, where participants are asked to count seconds in their minds while reading. This approach addresses a significant challenge in interval-based studies: the reliance on frequent participant feedback to collect ground-truth labels. Controlled distraction interventions, such as the one proposed in this study, are essential for understanding how distractions impact cognitive tasks like reading. Reading is a cognitively demanding activity that requires sustained attention, and even minor distractions can disrupt comprehension and focus. By simulating internal distractions in a controlled manner, these interventions allow researchers to systematically study the effects of distraction on reading behavior and cognitive performance [187].

To evaluate the impact of the counting distraction, we used a remote eye tracker (Tobii 4C) to capture participants' eye movements during reading tasks. Figure 1 illustrates two examples of scanpaths: one without the distraction condition (top) and one with the distraction condition (bottom). The results demonstrate that the counting distraction increases fixation duration and the frequency of losing the reading location, particularly during line breaks. These findings highlight the effectiveness of the proposed intervention in inducing measurable changes in reading behavior [187].

Such controlled interventions are particularly valuable as they provide a systematic way to study the effects of distraction while addressing key challenges in distraction research, such as reducing reliance on participant feedback through semi-automatic collection of ground-truth data.

The research questions associated with this work are as follows:

- **RQ1.** How can a novel distraction intervention be designed to reduce reliance on participant feedback during reading tasks?
- **RQ2.** How effective are the developed evaluation methods for detecting distraction and interest in reading, based on data collected in a controlled reading task?

This work provides insights into how controlled distractions influence reading behavior and offers a foundation for future studies on attention and cognitive load in reading tasks.

4.1.1 Data Collection

Participants

The study involved 10 participants, including both male and female individuals, aged between 20 and 30 years. All participants were recruited voluntarily and provided informed consent prior to the experiment. They were informed about the purpose of the study, the use of eye-tracking sensors, and their right to withdraw from the study at any time. Detailed information regarding data consent and experimental procedures was communicated to all participants before their participation [187].

Experimental Design

The experiment aimed to assess how a counting distraction influenced reading habits. Participants were tasked with reading six news articles, each containing three paragraphs, shown on a computer screen. A specialized software application was created to display one paragraph at a time in a predetermined sequence. While the sequence of articles varied for each participant, the order of paragraphs within each article remained unchanged. Participants switched between two conditions: reading while experiencing a counting distraction (mentally counting seconds) and reading without any distractions. The order of conditions was fixed to alternate (i.e., first with distraction, then without distraction). In total, each participant reviewed 18 paragraphs, consisting of nine under the distraction condition and nine under the no-distraction condition [187].

To ensure that participants were paying close attention, a comprehension question requiring a yes/no answer followed each paragraph. In addition, participants provided their subjective feedback for each paragraph by rating their interest and perceived difficulty on a six-point scale (0–100). For paragraphs that were read in the distraction condition, participants also indicated the time taken to complete the reading task. Reading behaviors were documented using a Tobii 4C remote eye tracker with a professional license. The data collected offers an in-depth look at participants' reading behaviors in both conditions [187].

4.1.2 Methodology

The methodology focuses on analyzing the effects of the counting distraction on reading behavior using the collected eye-tracking data. The study involved processing raw gaze data to extract meaningful features that reflect participants' reading patterns under both distraction and non-distraction conditions. These features were then used to evaluate the impact of the distraction intervention on cognitive and behavioral metrics. Upon visualizing the raw gaze data from the participants, it was observed that the gaze data of one participant were frequently missing due to a calibration issue. As a result, the data from this participant were excluded, and the gaze data of 9 participants were utilized for further machine learning tasks. The methodology is adapted based on our published work [187].

Data Preprocessing

The raw gaze data collected from the Tobii 4C eye tracker were preprocessed to ensure quality and consistency. To increase the number of training samples, a sliding window mechanism was incorporated during pre-processing. This approach allowed for the segmentation of gaze data into smaller, overlapping windows, enabling more robust feature extraction and model training [187].

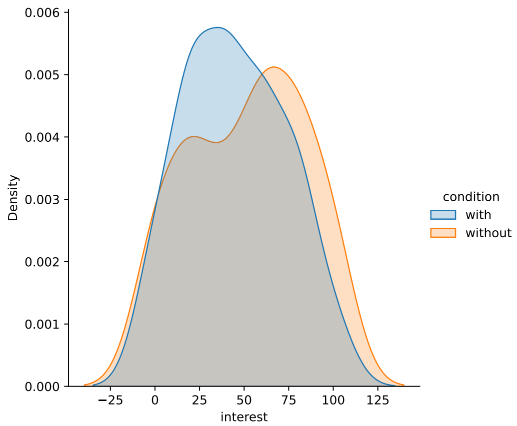
Feature Extraction

The features associated with fixations and saccades were derived from the processed gaze data. Saccades were also divided into three distinct categories: forward saccades (typical reading actions moving in the forward direction), backward saccades (returning to a previous line of text, identified by negative differences in horizontal

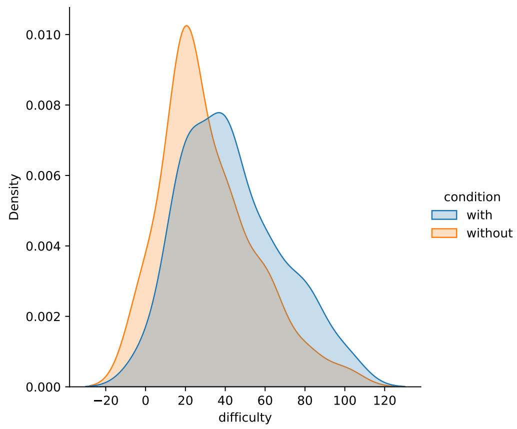
TABLE 4.1: The list of features

No.	Name
1	Fixation count
2, 3	Fixation duration {mean, std}
4, 5	Forward saccade length {mean, std}
6, 7	Forward saccade angle {mean, std}
8, 9	Backward saccade length {mean, std}
10, 11	Backward saccade angle {mean, std}
12, 13	Fixation duration {mean, std}
14, 15	Line breaks lengths {mean, std}
16, 17	Line breaks angles {mean, std}
18, 19	Pupil diameter {mean, std}

gaze), and line breaks (moving focus from one line to another). The extracted features are summarized in Table 4.1. These include fixation count, fixation duration (mean and standard deviation), saccade lengths and angles (mean and standard deviation for forward, backward, and line breaks), and pupil diameter (mean and standard deviation). These features were designed to capture key aspects of reading behavior under both distraction and non-distraction conditions [187].



(A) Interest Levels



(B) Difficulty Levels

FIGURE 2: Distributions of participants' subjective evaluations of texts under with and without counting conditions.

Model Training

The study evaluated the practicality of three machine learning techniques: Support Vector Machine (SVM), Random Forest (RF), and Multi-Layer Perceptron (MLP). These models aimed to predict two cognitive states: distraction and interest. The task was framed as a binary classification problem, utilizing the extracted features as inputs and assigning binary labels as outputs.

For the *distracted* label, paragraphs read under the counting distraction condition were labeled as 1 (distracted), while those read without distraction were labeled as

0. For the *interest* label, participants' interest ratings (on a scale of 0–100) were binarized using a threshold of 40, based on the distribution of interest levels (Figure 2a). Similarly, the difficulty levels of the texts, as evaluated by participants, are shown in Figure 2b. These distributions illustrate the subjective evaluations of texts under both counting and non-counting conditions, and the threshold effectively separated low-interest and high-interest paragraphs for classification [187].

Evaluation Protocol

The models were evaluated using three validation techniques to assess their performance under different scenarios:

- **User-independent (leave-one-participant-out):** The model was trained on the gaze data of all participants except one, with the excluded participant's data used as the validation set. This approach tested the model's ability to generalize to unseen participants.
- **User-dependent (leave-one-document-out for one participant):** For each participant, the model was trained on the gaze data of all but one document, with the remaining document used for validation. This approach evaluated the model's ability to predict distraction and interest for the same participant across different documents.
- **Document-independent (leave-one-document-out for all participants):** In this approach, the model was trained on the gaze data of all but one document (across all participants), with the excluded document used as the test set. This tested the model's ability to generalize across different documents.

The performance of the models was measured using precision, recall, F1-score, and accuracy. These metrics were chosen to balance the trade-off between false positives and false negatives, depending on the use case [187].

4.1.3 Results and Analysis

The performance of three classifiers—SVM, Random Forest, and MLP was evaluated using three validation methods: User-Independent (UI), User-Dependent (UD), and Document-Independent (DI). The results for both the *distraction* and *interest* classification tasks are summarized in Table 4.2. The results are based on our published work [187].

For distraction classification, SVM achieved its best performance under the DI configuration, with the highest precision (0.662). RF performed well under the UI configuration, achieving the highest recall (0.613), while MLP excelled under the DI configuration, attaining the highest F1-score (0.621) and accuracy (0.582). These

TABLE 4.2: Summary of the classification performance (UI: User-Independent; UD: User-Dependent; DI: Document-Independent)

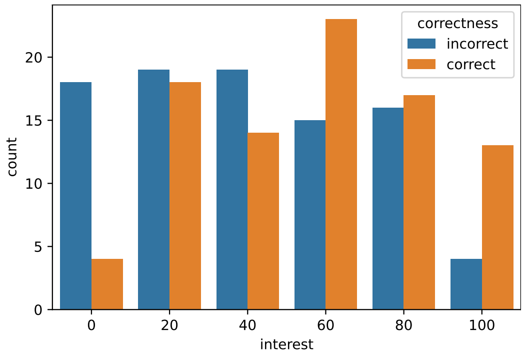
Model	Validation	Distraction				Interest			
		Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
SVM	UI	0.550	0.482	0.514	0.508	0.648	0.628	0.637	0.521
	UD	0.605	0.560	0.582	0.566	0.704	0.727	0.715	0.631
	DI	**0.662**	0.555	0.600	0.580	0.672	0.598	0.632	0.535
RF	UI	0.565	**0.613**	0.588	0.538	0.672	**0.898**	**0.769**	**0.638**
	UD	0.589	0.562	0.575	0.553	0.697	0.750	0.723	0.614
	DI	0.585	0.569	0.572	0.518	0.643	0.777	0.703	0.562
MLP	UI	0.551	0.558	0.554	0.517	0.671	0.781	0.722	0.590
	UD	0.561	0.535	0.548	0.524	**0.707**	0.721	0.714	0.612
	DI	0.641	0.611	**0.621**	**0.582**	0.659	0.707	0.681	0.558

results suggest MLP as a strong candidate for detecting distraction, showing particularly robust performance in balancing precision and recall. For interest classification, RF outperformed the other models, especially under the UI configuration, achieving the highest recall (0.898), F1-score (0.769), and accuracy (0.638). SVM also performed well under the UD configuration, with the highest precision (0.704) and recall (0.727). MLP, while competitive, achieved its best Precision (0.707) under the UD configuration. This indicates that RF is highly effective for identifying interest, particularly when generalizing across users [187].

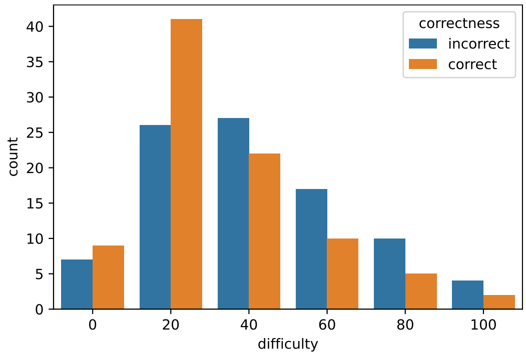
Overall, the results demonstrate that the choice of model and configuration significantly impacts performance. MLP is well-suited for the distraction task, while RF dominates the interest task. MLP, in particular, stands out for its balanced performance across both tasks, especially in the Distraction task under the DI configuration, where it achieved the highest F1-score and accuracy. This suggests that MLP is a versatile and reliable model, capable of adapting to different tasks and configurations with consistent results [187].

The overall distribution of difficulty and interest levels is depicted in Figure 2, which shows how the data is distributed under two conditions: with and without enforced distraction. In terms of interest levels, the distributions were marginally adjusted toward higher levels when distraction was absent, implying that participants found the texts more engaging when they were not distracted. Conversely, regarding difficulty levels, the distributions were somewhat adjusted toward higher levels when distraction was present, suggesting that participants viewed the texts as more challenging while distracted. These results emphasize the impact of distraction on participants' subjective perceptions of interest and difficulty [187].

The relationship between interest and difficulty levels with correctness is shown in Figure 3. As seen in Figure 3a, higher interest levels are generally associated with better correctness, indicating that participants performed better when they were more engaged with the text. On the other hand, Figure 3b shows that correctness decreases as the difficulty level increases, suggesting that participants struggled more



(A) Interest vs. Correctness



(B) Difficulty vs. Correctness

FIGURE 3: Relationship between objective comprehension (measured by the number of correct answers) and subjective difficulty and interest levels under the influence of the counting distractor.

with texts they perceived as difficult. These results emphasize the role of subjective difficulty and interest in shaping participants' objective comprehension [187].

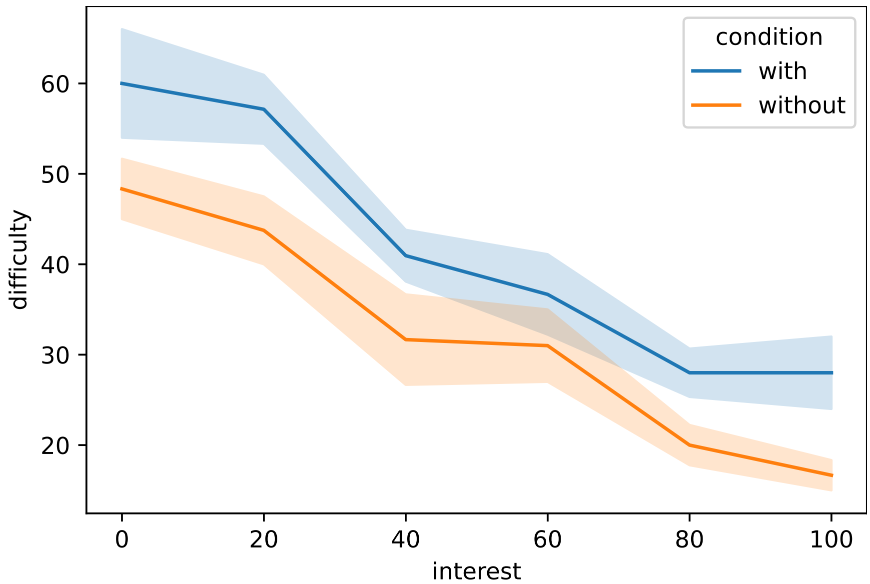


FIGURE 4: Relation with Distraction

Figure 4 explores how distraction affects the connection between interest and difficulty levels. As anticipated, a negative correlation was observed between interest and difficulty, indicating that texts viewed as more challenging were generally found to be less engaging. However, levels of difficulty were consistently higher when distraction was applied, independent of the interest level. This result implies that the counting distractor heightened the perceived difficulty of texts across all levels of interest, further illustrating the significant effect of distraction on participants' assessments [187].

4.1.4 Discussion

This study employed an eye-tracking approach to investigate the effects of distraction on cognitive processes during reading tasks. A counting distractor was introduced, requiring participants to count seconds in their minds while reading. This method streamlined the ground truth gathering process, which traditionally relies on self-reported surveys or questionnaires. By avoiding such subjective measures, the study minimized potential biases that could arise from participants' self-assessments. The discussion is based on our published work [187].

Impact of Distraction on Reading Behavior

The research investigated the levels of interest and difficulty of texts under two scenarios: with distractions and without. Findings showed that participants indicated greater interest when they were free from distractions. This implies that distractions increase the effect of text difficulty and lower participants' engagement with the material. These results are consistent with earlier studies, which have shown that distractions can directly hinder cognitive functions. For example, engaging in thoughts unrelated to the task often leads to divided attention [18], resulting in adverse outcomes such as challenges in maintaining focus, decline in mood [99], and weakened memory retention.

When subjects faced distractions, their difficulty ratings were higher, which adversely affected their accuracy on related comprehension questions. This finding supports the study's hypothesis, suggesting that a distracting element increases task difficulty by competing for limited attentional resources. Similar outcomes were noted by Navalpakkam et al. [130], who found that under significant distraction, participants invested more effort, had increased fixations, and often revisited sections they had already read. These outcomes underscore the considerable cognitive load that distractions create during reading activities [187].

Performance of Machine Learning Models

The performance of machine learning models for distraction and interest detection was evaluated using various validation techniques. Among the models, the Multi-Layer Perceptron (MLP) with a document-independent approach achieved the best accuracy and F1-score for distraction detection. For interest detection, the Random Forest model with a user-independent approach outperformed others in terms of accuracy, F1-score, and recall, as shown in Table 4.5 [187].

User-Dependent vs. User-Independent Validation

The study also evaluated user-dependent versus user-independent validation methods. The outcomes indicated that accuracy was marginally better across all models with user-dependent validation. These results do not provide enough evidence to definitively conclude that the differentiation between attention and distraction is significantly influenced by the user. Additional studies involving larger datasets and a more varied group of participants are necessary to investigate how individual differences may impact distraction detection [187].

4.1.5 Limitations and Future Work

While this study provides valuable insights into the effects of distraction on reading behavior and the performance of machine learning models for distraction and interest detection, it is not without limitations. One of the primary limitations is the relatively small dataset size. Although a sliding window algorithm was employed to increase the number of training samples, the dataset remained insufficient for robust evaluation of the machine learning models. This limitation likely impacted the generalizability of the results, particularly for distraction detection, where the performance metrics were less consistent compared to interest detection. Future studies should aim to collect larger datasets to improve the reliability and robustness of the models.

Another limitation lies in the use of a single type of distractor (i.e., the counting task). While this approach effectively induced distraction, it may not fully capture the diversity of real-world distractions, such as environmental noise, multitasking, or digital notifications. Future work could explore the effects of different types of distractors to better understand their impact on reading behavior and cognitive processes.

The study primarily focused on user-independent and user-dependent validation techniques but did not extensively investigate individual differences in attention and distraction. Factors such as cognitive capacity, reading proficiency, and prior knowledge may influence how participants respond to distraction. Future research could incorporate these individual differences to develop more personalized and adaptive models for distraction and interest detection.

Finally, while the machine learning models demonstrated promising results, particularly for interest detection, there is room for improvement in feature engineering and model design. Incorporating additional features, such as semantic or contextual information from the text, could enhance the models' ability to detect distraction and interest. Furthermore, exploring advanced deep learning architectures, such as attention-based models, may provide better performance and generalizability.

In summary, addressing these limitations in future work will not only improve the robustness of the findings but also contribute to the development of more effective systems for monitoring and mitigating distraction in real-world scenarios. Future work should focus on collecting larger datasets to improve model performance and exploring additional features, such as contextual or semantic information, to enhance the detection of distraction and interest. Investigating the role of individual differences in attention and distraction could provide deeper insights into user-specific patterns and improve the personalization of attention-monitoring systems [187].

4.1.6 Conclusion

This study introduced a novel approach to capture and model distraction during reading tasks using eye-tracking data. A semi-automated method for gathering ground truth was developed by employing a counting distractor, where participants were asked to count seconds in their minds while reading. This approach minimized reliance on traditional self-reported measures, providing a more objective framework for studying the effects of distraction on cognitive processes. The experiment analyzed the correlation between participants' subjective ratings of interest and difficulty under distracted and non-distracted conditions. The findings revealed that distraction significantly influenced participants' reading behavior, reducing their interest levels and increasing their perceived difficulty of the text. To predict distraction and interest levels, machine learning models were implemented using different cross-validation techniques. The results demonstrated that the MLP model with document-independent validation achieved the best performance for distraction detection, while the Random Forest model with user-independent validation outperformed others for interest detection. This research contributes to the growing body of work on understanding and modeling distraction during reading tasks. The proposed approach provides a foundation for developing systems that can monitor and mitigate distraction in real-world scenarios, such as educational or professional environments [187].

4.2 Cognitive Control Training for Digital Distraction Management

This section is adapted from our published work *"Digital Distractions in Reading: Investigating Impact of Cognitive Control Training on Reading Behavior and Outcomes"* [159]. The presence of digital notifications and pop-ups in modern reading environments poses a significant challenge to maintaining focused attention and achieving effective reading comprehension [35]. These interruptions disrupt the natural flow of reading, reduce comprehension, and increase frustration, making it essential to understand their impact on digital reading behaviors [118]. Notifications vary in frequency, placement, and content type, with some, such as reminders, eliciting quick acknowledgments, while others, like advertisements, demand prolonged attention. Distraction management remains a critical challenge, particularly in learning environments where digital devices are increasingly used [159].

This study investigates the effects of various types of distractions, including notifications, alerts, ads, and reminders, on reading comprehension, eye movements, and reader experience. Through an experiment involving 22 participants reading 10 documents under varied distraction conditions, we collected eye-tracking metrics and survey responses to assess the impact of these interruptions. The results revealed significant disruptions to reading flow, reduced comprehension, and heightened frustration levels among participants exposed to distractions. These findings underscore the need for strategies to mitigate the adverse effects of digital interruptions on reading performance [159].

Building on prior research [17, 136], we explored the potential of cognitive control training to enhance distraction management and improve reading outcomes. Participants were divided into two groups: a control group that received no training and an experimental group that underwent a 15-minute cognitive control training session before reading texts with varied notification types. The results demonstrated that cognitive control training improved reading comprehension and focus, even in environments with frequent distractions. This suggests that training in cognitive control can enhance individuals' ability to manage task-switching demands and maintain attention during reading [159].

The primary goal of this research is to understand natural reading behavior under the influence of distractions and to evaluate the role of cognitive control training in mitigating their effects. To achieve this, the study addresses the following research questions:

- How do varied distractions, such as notifications and pop-ups, impact reading behavior, comprehension, frustration levels, and document readability?
- What insights can be gained from eye-tracking data to better understand the relationship between distractions and reading performance?

- Can cognitive control training improve readers' ability to manage distractions and enhance comprehension in digital reading environments?

These findings have important implications for designing digital reading interfaces that minimize disruptions and promote a more seamless and productive reading experience.

4.2.1 Data Collection

The data collection process involved monitoring participants' reading behavior under controlled conditions while introducing distractions in the form of notifications and pop-ups. Eye-tracking data, subjective feedback, and comprehension scores were collected to analyze the impact of distractions and evaluate the effectiveness of cognitive control training. This section is based on the procedures detailed in our published work [159].

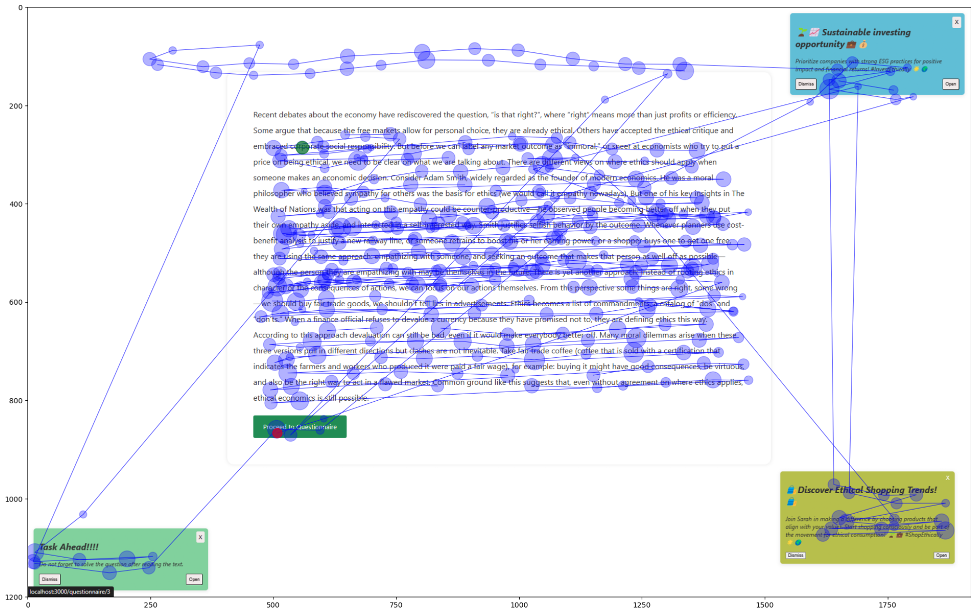
Participants

A total of 22 university students (14 female and 8 male, aged 23–32, M=26.48, SD=3.20) were recruited for this study. All participants provided informed consent before joining the experiment and were free to withdraw at any point. Detailed information about data consent, sensor usage, and the experiment sessions was communicated to all participants prior to their involvement [159].

Experimental Design

Participants were instructed to read ten texts, each approximately 500 words in length, covering general topics to ensure no prior knowledge advantage. Eye movements during reading were recorded using a Tobii Pro remote eye-tracker with a sampling frequency of 90 Hz. The texts were displayed on a desktop computer with the eye-tracker mounted on the screen in a controlled environment, ensuring consistent screen size and resolution for all participants.

The study employed a mixed-method approach to investigate the effect of cognitive control training (CCT). Eleven participants (experimental group) performed an adaptive version of the Paced Auditory Serial Addition Test (PASAT) before reading the texts, while the remaining 11 participants (control group) did not undergo any training. During the reading sessions, notifications and pop-ups, such as sales advertisements, emails, or reminders, appeared at random intervals and locations (e.g., the corner or top margin of the display). Participants were not instructed on whether to close the notifications, allowing for natural responses. To assess reading performance, participants completed both objective and subjective comprehension questionnaires after reading each text. Additionally, document-wise subjective measures of distraction, comprehension, frustration, and overall readability were collected on a 5-point scale, with 1 being the lowest and 5 the highest value [159].



(A) Highly distracted participant (no cognitive control training)

(B) Less distracted participant (with cognitive control training)

FIGURE 5: Comparison of gaze scanpaths between (a) a highly distracted participant who did not receive cognitive control training and (b) a less distracted participant who received cognitive control training. Both participants viewed the same document.

4.2.2 Methodology

The methodology of this study involved recording participants' eye gaze data during reading sessions to assess their comprehension, distraction, frustration, and overall document readability. A Tobii Pro remote eye-tracker with a pro license was used to capture gaze coordinates and pupil diameters at a sampling rate of 90 Hz, providing robust data for analysis. The methodology is based on our published work [159].

Data Preprocessing

The raw gaze data collected from the eye-tracker was preprocessed to identify fixations and saccades, which were then used to generate document scanpaths. The x and y gaze coordinates, along with left and right pupil diameters, served as primary features for predicting distraction, comprehension, and frustration levels. A sliding window approach with a length of 10 seconds and 50% overlap was applied to segment the data for analysis. Figure 5 illustrates a comparison of gaze data from two participants reading the same document, one of whom without cognitive control training (Figure 5a) while the other who underwent the training (Figure 5b) [159].

Models

To analyze and classify participants' distraction, comprehension, and frustration levels, three neural network models were employed: CNN-LSTM, InceptionTime, and Transformer. The CNN-LSTM model combined convolutional neural networks (CNNs) for extracting spatial features with long short-term memory (LSTM) networks to capture temporal dependencies in the gaze data. This hybrid approach allowed the model to effectively classify reading patterns based on both spatial and temporal characteristics. InceptionTime, on the other hand, utilized an ensemble of inception blocks, where each block consisted of convolutional layers designed to extract multiscale temporal features from the gaze data. This model was particularly effective in identifying patterns across varying time scales. Lastly, the Transformer model leveraged self-attention mechanisms to identify dependencies and relationships within the gaze data, enabling a deeper understanding of the participants' reading behavior. Together, these models provided a comprehensive framework for analyzing the complex dynamics of gaze data [159].

Classification

The neural networks were designed to predict three primary outcomes: distraction, comprehension, and frustration. Participants rated their experiences regarding each outcome using a 5-point Likert scale, with 1 indicating 'very low' and 5 indicating 'very high'. The self-reported ratings were subsequently classified into binary categories to simplify the classification process. Ratings from 1 to 3 were combined as 'Low', whereas ratings of 4 and 5 were merged as 'High'. This binary classification method enabled the models to differentiate between low and high levels of distraction, comprehension, and frustration, offering insights into how participants reacted to digital distractions while reading [159].

Evaluation Protocol

To assess the models' performance, two techniques for data partitioning were utilized: participant-independent and document-independent methods. The participant-independent method, commonly known as leave-one-participant-out (LOPO), involved selecting one participant's data as the test set while utilizing the data from the other participants to train the model. This procedure was conducted repeatedly for each participant, and the overall model performance was evaluated by calculating the average accuracy across all iterations. Similarly, the document-independent approach, called leave-one-document-out (LODO), involved using one document as the test set while the rest of the documents served to train the model. This process was also repeated for each document, with the average accuracy computed across all iterations. These evaluation protocols ensured that the models were evaluated on previously unseen participants and documents, thereby providing a thorough assessment of their generalizability and efficacy in real-world applications [159].

TABLE 4.3: Comparison of Models with LOPO and LODO validation methods

Models	Validation	Distraction		Comprehension		Frustration	
		Acc	F1	Acc	F1	Acc	F1
CNN-LSTM	LOPO	0.72	0.67	0.75	0.67	0.79	0.70
	LODO	0.79	0.71	0.80	0.72	0.81	0.74
InceptionTime	LOPO	0.70	0.65	0.72	0.63	0.75	0.69
	LODO	0.73	0.68	0.77	0.69	0.79	0.71
Transformers	LOPO	0.74	0.68	0.76	0.66	0.79	0.72
	LODO	**0.80**	**0.73**	**0.83**	**0.75**	**0.85**	**0.77**

4.2.3 Results and Analysis

This study utilized binary classification to predict participants' distraction, comprehension, and frustration levels using three deep neural network models: CNN-LSTM, InceptionTime, and Transformer. The performance of these models was evaluated using two validation methods: leave-one-participant-out (LOPO) and leave-one-document-out (LODO). The results, summarized in Table 4.3, indicate that the models performed better overall with the LODO approach, suggesting that the models generalized more effectively when tested on unseen documents compared to unseen participants. Among the three models, the Transformer consistently outperformed the others across all labels, demonstrating its ability to capture complex dependencies in the gaze data. The results and analysis are based on our published work [159].

Specifically, the Transformer model achieved the highest accuracy and F1-scores for all three outcomes. For distraction, it achieved an accuracy of 0.80 and an F1-score of 0.73. For comprehension, the model achieved an accuracy of 0.83 and an F1-score of 0.75. Similarly, for frustration, it achieved an accuracy of 0.85 and an

F1-score of 0.77. These results highlight the effectiveness of the Transformer model in identifying patterns in gaze data that are indicative of participants' cognitive and emotional states during reading tasks [159].

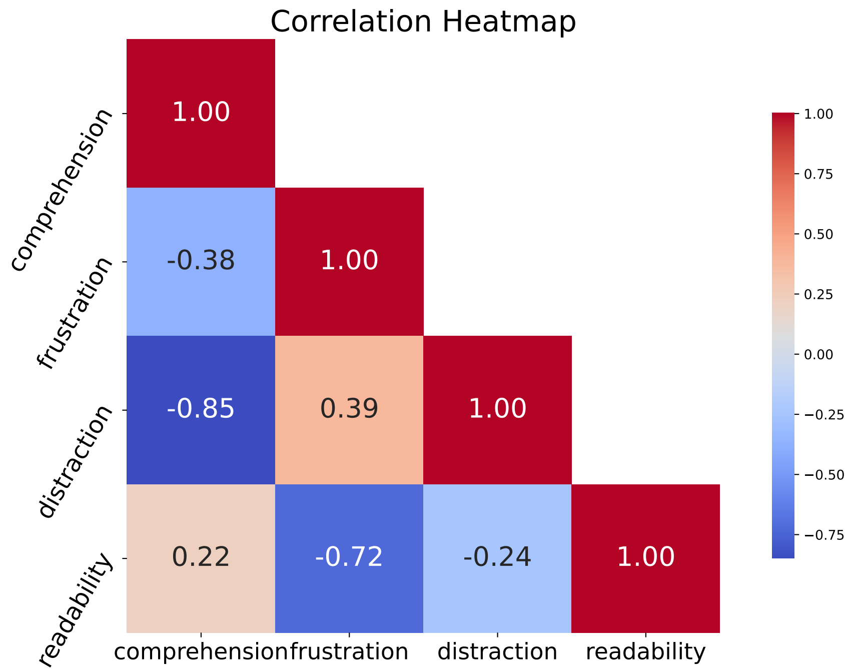


FIGURE 6: The correlation heatmap of the survey responses

In addition to model performance, a correlation analysis was conducted to explore the relationships between the key variables: distraction, comprehension, frustration, and readability. The correlation plot in Figure 6 revealed several significant relationships. A strong negative correlation was observed between comprehension and distraction (r = -0.85), indicating that as distraction levels increase, comprehension decreases. This finding aligns with expectations, as higher levels of distraction disrupt the reading flow and impair the ability to process and retain information. Conversely, lower distraction levels are associated with better comprehension, emphasizing the importance of minimizing interruptions in reading environments [159].

A moderate positive correlation was observed between frustration and distraction (r = 0.39), suggesting that as participants experience more distractions, their frustration levels also tend to rise. This relationship underscores the emotional toll that frequent interruptions can have on readers, further impacting their overall reading experience. Interestingly, a weak positive correlation was found between comprehension and readability (r = 0.22). While this indicates that higher comprehension is generally associated with better readability, the relationship is not as strong as the other correlations, suggesting that readability may be influenced by additional factors beyond comprehension alone. These findings have important implications for the design of digital reading environments and learning materials. The results highlight the need to minimize distractions and frustration to promote better comprehension and readability. By addressing these factors, designers and educators can create more effective and user-friendly reading interfaces that support focused and productive reading experiences [159].

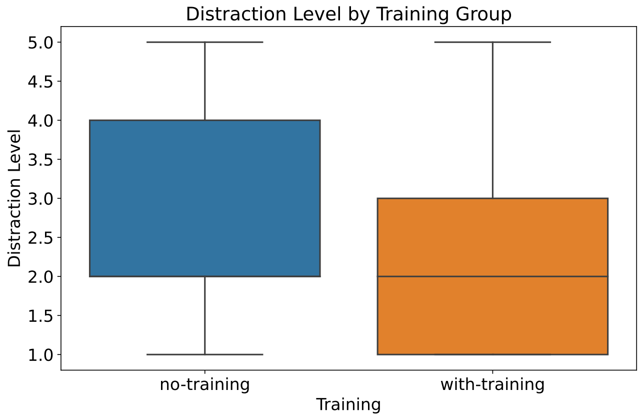
4.2.4 Discussion

The findings of this study highlight the significant impact of cognitive control training as an intervention on improving reading comprehension and managing distractions in digital reading environments. While the study also demonstrates the effectiveness of deep neural networks in predicting cognitive and emotional states, the primary focus remains on the role of cognitive control training in mitigating the adverse effects of distractions. Distractions, such as pop-ups and notifications, were shown to hinder comprehension and document readability while increasing frustration levels, as observed in both objective and subjective measures. The discussion is based on our published work [159].

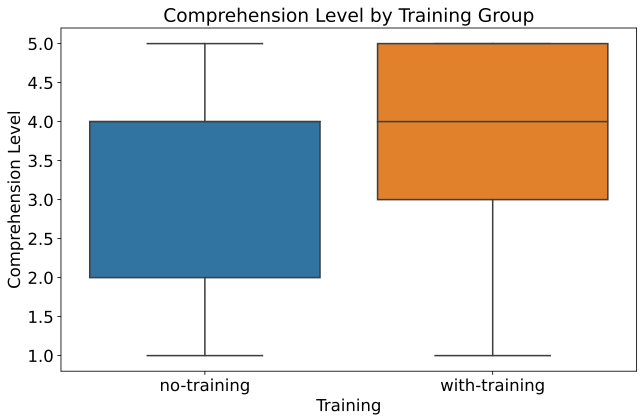
To investigate the impact of cognitive control training, participants were divided into two groups: one group read documents without any prior training (*P01-P11*), while the other group underwent cognitive control training (*P12-P22*) before reading. The results clearly indicate that participants who received training exhibited significantly improved comprehension and reduced distraction levels compared to the untrained group. This supports the hypothesis that cognitive control training can enhance an individual's ability to manage distractions and maintain focus during reading tasks [159].

The analysis of both objective and subjective questionnaires further validated these findings. As shown in Figure 7, the trained group demonstrated a notable decrease in distraction levels (Figure 7a) and an increase in comprehension levels (Figure 7b). Statistical analysis revealed that the training had a significant positive effect on comprehension (*p-value = 0.0059*) and distraction levels (*p-value = 0.021*). However, the training did not have a significant impact on frustration levels (*p-value = 0.899*), suggesting that while cognitive control training helps participants focus better and comprehend more effectively, it may not directly alleviate the emotional response to distractions. Figure 8 further illustrates the average ratings for comprehension, frustration, readability, and distraction across documents, showing that documents with higher comprehension and readability scores were generally associated with lower distraction and frustration levels [159].

These findings have important implications for the design of personalized learning systems and digital reading environments. By incorporating cognitive control training into educational tools, it may be possible to help users better manage distractions and improve their reading outcomes. This is particularly relevant in today's digital age, where frequent interruptions from notifications and pop-ups are



(A) Distraction ratings



(B) Comprehension ratings

FIGURE 7: Comparison of (a) distraction and (b) comprehension ratings for participants with and without cognitive control training.

common. The results suggest that cognitive control training could serve as a valuable intervention for enhancing focus and comprehension in such environments.

4.2.5 Limitations and Future Work

While this study provides valuable insights into the impact of cognitive control training on reading behavior, there are several limitations that should be addressed in future research. First, the small sample size of 22 participants limits the generalizability of the findings. Recruiting a larger and more diverse sample in future studies would help validate these results across different populations and contexts. Additionally, the study primarily relied on self-reported measures of comprehension, frustration, and readability, which may be subject to biases. Incorporating more objective measures, such as task performance metrics or physiological data (e.g., heart rate variability or skin conductance), could provide a more comprehensive understanding of the effects of cognitive control training.

Another limitation is the lack of control for individual differences in cognitive abilities, reading habits, and prior exposure to digital distractions. These factors may have influenced participants' responses and outcomes, and future studies should

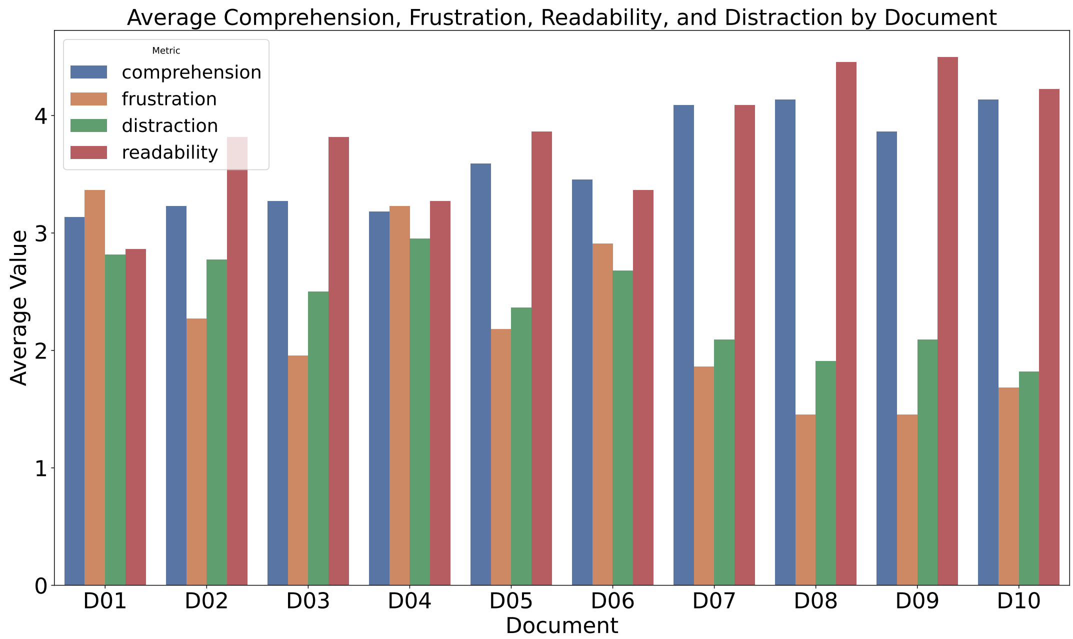


FIGURE 8: The average comprehension, frustration, readability, and distraction across each document as reported by the participants

consider more rigorous experimental designs to account for such variables. For example, stratifying participants based on their baseline cognitive control abilities or prior experience with digital reading environments could provide deeper insights into the effectiveness of the training.

Future work could also explore the long-term effects of cognitive control training. This study focused on the immediate impact of a single training session, but it remains unclear whether the benefits of training persist over time or require repeated sessions to maintain effectiveness. Additionally, expanding the scope of the study to include different types of reading tasks, such as academic or technical reading, could help determine whether the training is equally effective across various contexts. Finally, integrating cognitive control training into real-world digital reading platforms and evaluating its impact in naturalistic settings would provide practical insights for its implementation in educational and professional environments.

In summary, while this study demonstrates the potential of cognitive control training to improve reading outcomes and manage distractions, addressing these limitations and exploring new directions will be essential for advancing this research and its applications [159].

4.2.6 Conclusion

This study highlights the disruptive impact of notifications and pop-ups on reading comprehension, frustration levels, and the overall reading experience. Through an experiment involving 22 university students reading ten documents under varying distraction conditions, the findings demonstrate that frequent interruptions significantly hinder comprehension and increase frustration. The use of deep neural networks, including CNN-LSTM, InceptionTime, and Transformer models, proved effective in predicting distraction, comprehension, and frustration levels based on eye-tracking data, offering valuable insights into reading behavior in digital environments. A key contribution of this research is the evidence supporting the effectiveness of cognitive control training in mitigating the negative effects of distractions. Participants who underwent cognitive control training exhibited significantly improved comprehension and reduced distraction levels compared to those who did not receive training. These results underscore the potential of cognitive control training as a practical intervention to enhance focus and reading performance in environments prone to frequent interruptions. The findings also emphasize the need for effective notification and pop-up management strategies to minimize disruptions and promote optimal reading experiences. This study provides a foundation for future research into personalized interventions and adaptive systems that can help users navigate the challenges of digital reading in increasingly distraction-heavy environments. This research demonstrates the importance of both cognitive control training and thoughtful interface design in improving reading outcomes. By combining these approaches, it is possible to create digital reading environments that foster focus, comprehension, and a more seamless reading experience [159].

4.3 Emotion and Engagement Interventions

This section is adapted from our published work titled *"Toward an Interactive Reading Experience: Deep Learning Insights and Visual Narratives of Engagement and Emotion"* [160] and *"Real-Time Feedback on Reader's Engagement and Emotion Estimated by Eye-Tracking and Physiological Sensing"* [138]. In the context of personalized interventions, understanding and addressing user engagement and emotional states is critical for creating effective and adaptive systems [200]. Engagement and emotion are key factors that influence the quality of user interaction, particularly in e-learning environments where sustained focus and motivation are essential for success [38]. This section explores an innovative approach to monitoring and enhancing user engagement and emotional states through a personalized, interactive dashboard. By leveraging predictive models and user-dependent approaches, the system provides feedback and interventions to optimize the user's experience. The goal is to ensure a deeper connection between the user and the task, ensuring both emotional well-being and sustained engagement.

The proposed system integrates advanced deep learning models to predict engagement levels and emotional states, presenting the results through an intuitive user interface. The dashboard not only visualizes these metrics but also incorporates an alert and motivational system to actively guide users toward maintaining optimal engagement. This section builds upon the data collection, methodology and model evaluation processes discussed in Section 3.4 [138, 160].

4.3.1 Architecture

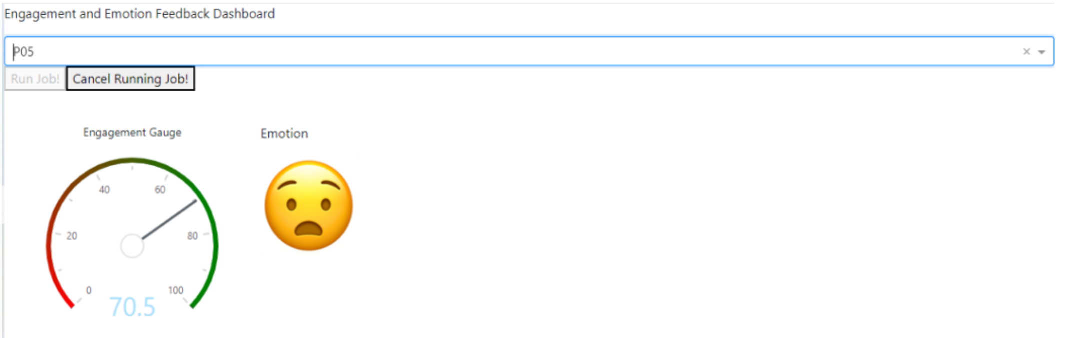
The architecture of the proposed system is designed to seamlessly integrate sensor data, model prediction, and user feedback. It consists of the following key components [138, 160]:

- **Data Module:** This module gathers sensor data from users, such as physiological signals and gaze data. These inputs are essential for predicting engagement and emotional states.
- **Prediction Engine:** At the core of the system lies a deep learning-based prediction engine. This engine processes the data to estimate two key metrics:
  - *Engagement Level:* Predicted as a continuous value ranging from 0 to 100, representing the user's focus and involvement in the task.
  - *Emotional State:* Predicted using a two-dimensional model of emotion (valence and arousal), which categorizes emotions such as happiness, frustration, or boredom.
- **Feedback and Alert System:** The system generates real-time feedback based on the predictions. Alerts are triggered when engagement levels drop below a

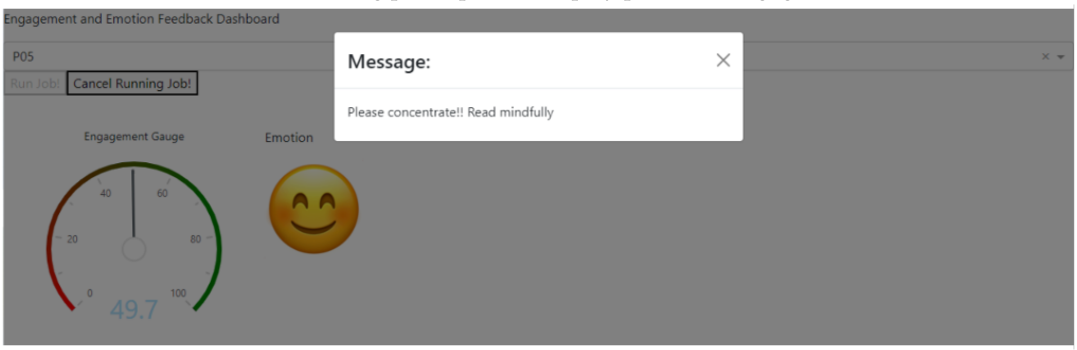
threshold, while motivational messages are displayed during periods of high engagement.

- **User Interface:** The dashboard serves as the primary medium for user interaction, providing visual and textual feedback to enhance the user's experience.

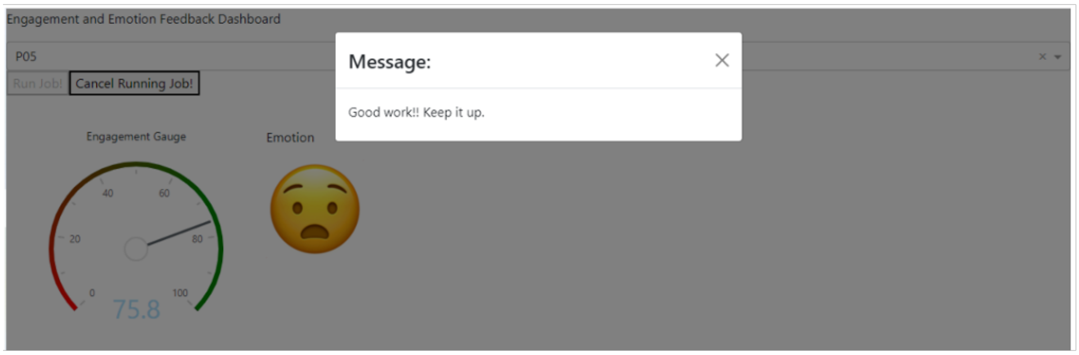
This modular architecture ensures scalability and adaptability, allowing the system to be tailored to individual users and specific tasks.



(A) Dashboard after selecting participant to display predicted engagement and emotion



(B) Dashboard notifying the participant to concentrate on text as it detected a fall in engagement level



(C) Dashboard motivating the participant to continue doing the good work since the engagement levels are high over a consistent period

FIGURE 9: Visualizations of the engagement monitoring dashboard: (a)displaying predicted engagement and emotion for a selected participant, (b) notifying low engagement, and (c) motivating high engagement.

4.3.2 Model Selection and Integration

The effectiveness of the system relies on the selection and integration of predictive models. A user-dependent approach was adopted to ensure personalized predictions, as engagement and emotional responses can vary significantly between individuals. The following steps were taken to identify and integrate the best-performing models:

- **Model Selection:** Various deep learning models were evaluated for their ability to predict engagement and emotional states. The models were trained and tested on user-specific datasets to account for individual differences. Metrics such as accuracy, precision, and recall were used to identify the most reliable models.
- **Integration:** The selected models were integrated into the system, enabling real-time predictions. The engagement model outputs a probability distribution over engagement levels, which is then mapped to a 0–100 scale for visualization. Similarly, the emotion model predicts valence and arousal values, which are used to generate an appropriate emotion emoji.
- **Threshold Calibration:** Thresholds for engagement alerts and motivational messages were calibrated based on user feedback and experimental results. This ensures that the system's interventions are both timely and meaningful.

By prioritizing personalization and accuracy, the system delivers highly relevant feedback to users, enhancing their overall experience [138, 160].

4.3.3 Interactive User Interface Design

The user interface (UI) plays a pivotal role in the success of the intervention, which is shown in Figure 9. The dashboard was designed to be intuitive, engaging, and informative, with the following features [138, 160]:

- **Engagement Gauge:** A visual gauge displays the user's engagement level on a scale of 0 to 100. This provides a clear and immediate representation of the user's focus and involvement.
- **Emotion Emoji:** A dynamic emoji reflects the user's emotional state, based on the predicted valence and arousal values. This feature adds a humanized and relatable element to the feedback.
- **Alert System:** When engagement levels drop below a predefined threshold, the dashboard displays an alert message, prompting the user to refocus. For example, Figure 9b illustrates an alert notifying the user to concentrate on the text.

- **Motivational Messages:** During periods of high engagement, the dashboard displays motivational messages to reinforce positive behavior. Figure 9c demonstrates a motivational message encouraging the user to maintain their elevated engagement levels.
- **Customizability:** Users can adjust the dashboard settings, such as the thresholds for alerts and motivational messages, to suit their preferences and needs.

The UI design prioritizes simplicity and clarity, ensuring that users can easily interpret the feedback and take appropriate actions.

4.3.4 Discussion and Limitations

The proposed system represents a significant advancement in personalized interventions for engagement and emotion. By combining predictive modeling with an interactive dashboard, the system addresses key challenges in maintaining user focus and motivation. This approach has the potential to transform how engagement is monitored, particularly in e-learning environments where maintaining student focus is a persistent challenge. By providing real-time feedback and interventions, the dashboard can enhance learning outcomes and reduce dropout rates.

In professional settings, the system could be used to monitor and improve employee focus during tasks, leading to increased productivity and job satisfaction. Additionally, the emotion prediction feature could be extended to identify signs of stress or frustration, enabling timely interventions to support mental well-being. These applications highlight the versatility of the system in addressing engagement and emotional challenges across various domains.

However, there are limitations to consider. The accuracy of the predictions depends heavily on the quality of the input data, which may vary across users and contexts. For instance, physiological or behavioral signals used for prediction can be influenced by external factors such as environmental noise or user fatigue, potentially reducing the reliability of the system. Furthermore, the user-dependent approach, while effective in personalization, requires sufficient data for each individual, which may not always be feasible in real-world scenarios.

Real-time monitoring, while a key goal of the system, presents additional challenges. Latency, real-world distractions, and user interruptions can hinder the system's ability to provide accurate and timely insights. Achieving true real-time feedback may require seamless integration into users' routines, which could be difficult to implement in diverse and unpredictable environments [160].

The effectiveness of the alert messages also warrants further investigation. While the system provides customized alerts to notify users of engagement drops or to motivate them during high engagement periods, the impact of these messages on user behavior and emotional regulation remains uncertain. User responses to alerts may vary significantly, and their ability to understand and act on these messages should be carefully assessed in future studies [160].

Despite these limitations, the system provides a promising foundation for personalized interventions in engagement and emotion. Future research will focus on refining the prediction models, addressing real-time monitoring challenges, improving the effectiveness of alert messages, and exploring its applications in diverse real-world settings. By addressing these challenges, the system can be further optimized to enhance user experiences and outcomes across various domains [138, 160].

4.3.5 Conclusion

This section presented a novel system for monitoring and enhancing user engagement and emotional states through personalized interventions. By leveraging advanced predictive models and an interactive dashboard, the system provides feedback that is both intuitive and actionable. The architecture of the system integrates data collection, machine learning-based predictions, and user feedback mechanisms to create a seamless and adaptive experience. The use of a user-dependent approach ensures that the predictions are tailored to individual users, offering a high degree of personalization that is critical for accurately capturing fluctuations in engagement and emotional states.

The interactive dashboard serves as the integral part of the system, offering users a clear and engaging way to visualize their engagement levels and emotional states. Features such as the engagement gauge, emotion emoji, and alert system provide immediate feedback, helping users to better understand their cognitive and emotional responses during tasks. The motivational messages further enhance the user experience by reinforcing positive behaviors and encouraging sustained engagement. These features collectively make the system not only a tool for monitoring but also an active participant in improving user focus and emotional well-being. The system's potential applications extend across various domains, including e-learning, workplace productivity, and mental health. In e-learning environments, the ability to provide personalized feedback and maintain high engagement levels can significantly improve learning outcomes and reduce dropout rates. This personalized and interactive approach represents a meaningful step in the design of systems that prioritize user engagement and emotional health, offering a foundation for more adaptive and user-centered technologies in the future [138, 160].

4.4 Real-Time Gaze Based Adaptive Learning System

This section and the related contents are adapted from our published work titled *"Gaze-Driven Adaptive Learning System With ChatGPT-Generated Summaries"* [156]. In the rapidly changing field of educational technology, the primary objective remains to improve student involvement and educational outcomes [141, 170]. Both engagement and comprehension are essential elements of successful learning, directly impacting a learner's capacity to grasp, remember, and utilize knowledge [23, 69]. Although online learning platforms have transformed access to educational resources, they also introduce challenges like limited customization, lower engagement, and inadequate feedback. Tackling these issues requires the creation of adaptive learning systems that can utilize real-time data to deliver personalized support and interventions [156].

Recent developments in artificial intelligence and real-time data analysis present promising prospects for establishing more responsive and adaptable learning environments [197]. Adaptive e-learning systems aim to customize educational materials to fit the unique needs, preferences, and learning styles of individual students [132, 173]. However, numerous current systems depend on fixed models of learner behavior, which do not adequately reflect the dynamic and intricate nature of human learning. Moreover, these systems frequently lack the capability to deliver real-time feedback, which is crucial for sustaining engagement and motivation [3, 20]. Engagement itself is a complex concept influenced by both emotional and cognitive states, and the capacity to monitor these states in real time can provide valuable insights into the learning experience. Such insights allow for timely interventions that can help re-engage learners and redirect their focus when necessary [156].

Even with these developments, creating efficient systems for real-time monitoring and intervention in engagement poses considerable challenges [75]. Conventional approaches often fail to effectively identify and react to the immediate emotional and cognitive states of students, especially during times of disengagement. Additionally, incorporating real-time analytics into educational models necessitates tackling issues surrounding privacy, ethics, and technological infrastructure. Although adaptive systems have shown promise in enhancing learning outcomes, their constraints in precisely evaluating and addressing fluctuating levels of engagement underscore the necessity for more innovative strategies [156].

To tackle these challenges, this study presents an innovative adaptive learning system driven by gaze. In contrast to prior methods that depend on delayed or indirect indicators of engagement, this system employs real-time gaze tracking as a continuous and non-intrusive approach to monitor student involvement. By examining gaze behaviors, the system is able to identify subtle fluctuations in engagement and offer prompt, customized interventions. These interventions are provided in the form of summaries generated by ChatGPT, aimed at re-engaging students and improving their understanding of the content [156].

The system utilizes real-time data to adaptively cater to the distinct requirements of every learner. When signs of disengagement are identified, the system takes action by producing brief, context-sensitive summaries of the educational content, which aids in redirecting the learner's focus and enhancing their comprehension. This study investigates two primary questions [156]:

- **RQ1.** How well can real-time gaze tracking data be employed to assess and monitor student engagement levels within adaptive learning environments?
- **RQ2.** To what degree do AI-generated adaptive interventions, initiated by real-time predictions of engagement, enhance student learning outcomes and understanding?

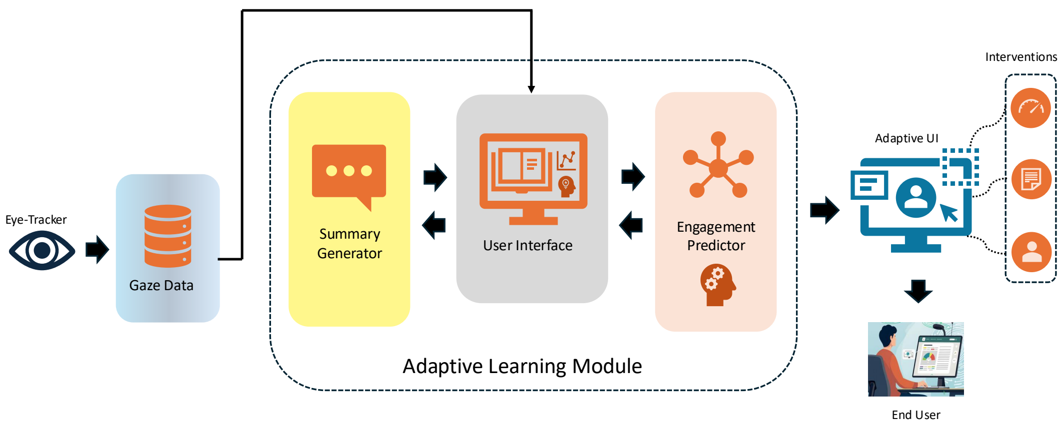


FIGURE 10: Architecture of the gaze-driven adaptive learning system, illustrating its key components

The system architecture, depicted in Figure 10, features a user interface that tracks eye movements and analyzes gaze data to assess levels of engagement. When signs of disengagement are identified, the system employs a ChatGPT-based API to create summaries of the learning content, which are presented to the learner in real time. This method addresses significant shortcomings of current adaptive learning systems, such as the absence of immediate intervention abilities and the failure to recognize subtle fluctuations in engagement. Through a controlled experimental study, this research establishes the effectiveness of the proposed system in enhancing learning outcomes. Participants who received adaptive interventions demonstrated improved comprehension and maintained higher engagement levels compared to those in a control group. Furthermore, the system was noted to reduce visual strain while keeping learners focused. These results highlight the promise of real-time adaptive interventions in enriching the educational experience and set the stage for future developments in adaptive e-learning technologies [156].

The contributions of this study are outlined as follows [156]:

- Creation of an innovative adaptive learning framework that combines real-time gaze-based engagement predictions with AI-generated text summaries.
- Comprehensive assessment of the system's efficacy in enhancing student engagement and comprehension, utilizing gaze tracking and survey data.
- Insights regarding the connection between cognitive processing, visual exertion, and interventions in adaptive learning.

By tackling the issues of real-time engagement assessment and support, this study pushes forward the field of personalized learning and showcases the potential of AI-enhanced adaptive systems to revolutionize educational experiences [156].

4.4.1 User Study and Data Collection

A user study was conducted to evaluate the effectiveness of the adaptive learning system in enhancing learners' comprehension, engagement, and self-confidence within a controlled learning environment. The study aimed to explore whether the system, which integrates gaze tracking and ChatGPT-generated summaries, could improve learners' understanding and retention of material, increase their engagement and motivation, and boost their confidence in their ability to learn. The user study and data collection process are based on our published work [156].

Participants

The research involved 22 university students (11 male and 11 female) aged between 22 and 29 years (M = 25, SD = 2.5). All participants were selected based on their high proficiency in English, verified through standardized language assessments and self-reported language usage, to ensure that language skills did not affect their understanding or reading experience. The study received approval from the DFKI Ethics Committee and was carried out following its ethical guidelines. Participants gave informed consent before taking part in the study and had the option to withdraw at any time [156].

Experimental Design

The research utilized a between-subjects design, with participants randomly allocated to either an experimental group (n=11) or a control group (n=11). The experimental group was provided with real-time engagement forecasts and summaries generated by ChatGPT, whereas the control group did not receive any adaptive measures. Participants were instructed to read a curated set of ten documents, each chosen to provoke different levels of engagement and comprehension. During the reading activity, gaze patterns and eye movements were captured using a Tobii 4C eye tracker with a professional license, functioning at a sampling rate of 90 Hz and attached to a display monitor. This configuration allowed accurate monitoring of participants' visual attention and engagement levels.

Participants took a quiz comprising four comprehension-based questions to evaluate their objective understanding of the material after reading each document. In addition, a post-survey was conducted to gather self-reported assessments of comprehension, engagement, and self-confidence regarding their responses to the quiz questions. The survey made use of a 7-point Likert scale, ranging from 1 (lowest level) to 7 (highest level), allowing participants to consider their perceived understanding, interest, and confidence. The questions from the survey are outlined in Table 4.4. By merging the objective comprehension scores from the quizzes with the subjective self-assessments, the research sought to deliver a thorough assessment of the system's influence. This combined method reduced possible biases inherent in self-reported data and provided a better understanding of the variations in comprehension, engagement, and confidence levels between the experimental and control groups [156].

TABLE 4.4: Post-Survey for Experiment Participants: Questions 1-3 were administered to all participants, while Questions 4-6 were specific to the experimental group.

No	Question	Type	Scale
1.	Rate the overall intensity of your engagement while reading the document.	Engagement	1-7
2.	Rate the overall level of comprehension or understanding while reading the document.	Comprehension	1-7
3.	Rate your self-confidence while answering the comprehension-based questionnaire.	Self-Confidence	1-7
4.	How well do you think your gaze behavior predicted by the system reflects your level of interest or engagement with the document?	Gaze-Reflect	1-7
5.	Rate the helpfulness of the summary in capturing the main points of the document.	Summary-Help	1-7
6.	How often did you refer back to the summary while reading?	Summary-Refer	1-7

4.4.2 Methodology

This research utilized a mixed-methods approach to develop and assess a gaze-driven adaptive learning system that incorporates LLM-generated summaries. The system was aimed at delivering personalized learning experiences by responding to each learner's unique affective states. To accomplish this, eye-tracking was combined with a ChatGPT-based summary generation system, allowing for real-time interventions and assistance for learners. This subsection outlines the methodology employed to design and develop the system, covering the system architecture, engagement prediction model, integration of the ChatGPT API, gaze feature metrics, and the training and evaluation of the model. The methodology and its associated subsections are based on our published work [156].

System Architecture

The adaptive learning system consists of four main components: a gaze tracking module, an engagement prediction model, a ChatGPT API integration, and a user interface. The system flow is as follows [156]:

Gaze Tracking Module The gaze tracking module gathers data from learners as they use the system. This information consists of the x and y coordinates of the learner's gaze point, along with additional data such as timestamps and pupil size. The raw gaze information is analyzed to extract fixation data, which indicates the points where the learner's gaze remains stable for a specified duration. This analysis includes filtering out noise and determining fixation points based on gaze behavior. The fixation data, which includes both gaze coordinates and pupil diameter measurements, is stored in an array and continuously updated in real time. The system utilizes raw gaze data to develop a real-time visual depiction of the user's gaze, while the fixation data serves to predict engagement levels, offering a deeper understanding of learner behavior and focus. Fixation data is sent at 5-second intervals, enabling the system to timely assess the user's engagement level and react accordingly [156].

Engagement Prediction Model The engagement prediction model is a deep learning algorithm that has been initially trained using data from prior studies related to interest and engagement detection [157]. This model utilizes fixation data and pupil diameter measurements as its inputs and generates an engagement score as the output. It has been trained on a collection of labeled instances, each consisting of fixation points, pupil diameter statistics, and an associated level of engagement. Engagement scores are predicted as binary outcomes, with '0' indicating low engagement and '1' signifying high engagement. The computed engagement score is transmitted to the user interface, where it is presented on an interactive dashboard. This dashboard provides real-time feedback to learners regarding their engagement levels, allowing them to monitor and adjust their focus as needed [156].

ChatGPT-Based Summary Generation To create summaries of the learning materials, the system utilizes an API based on ChatGPT. The gpt-3.5-turbo model from OpenAI was chosen for its capability to generate high-quality text outputs quickly. When summarizing a paragraph, the API is triggered with a structured prompt that centers on the given content. Summaries are capped at 80 tokens to maintain conciseness and informativeness. The temperature parameter is adjusted to 0.5, striking a balance between creativity and clarity. The produced summaries go through a refining process to eliminate unnecessary elements (such as URLs) and to ensure they are comprehensive. The final summary is chosen based on its coherence and relevance and is presented to the learner when their engagement levels are low. The summary generation process can be mathematically represented as follows [156]:

$$S_{\text{final}} = \text{Process}(S_i) \text{ where } S_i \in \{S_1, S_2, \ldots, S_n\} \tag{4.1}$$

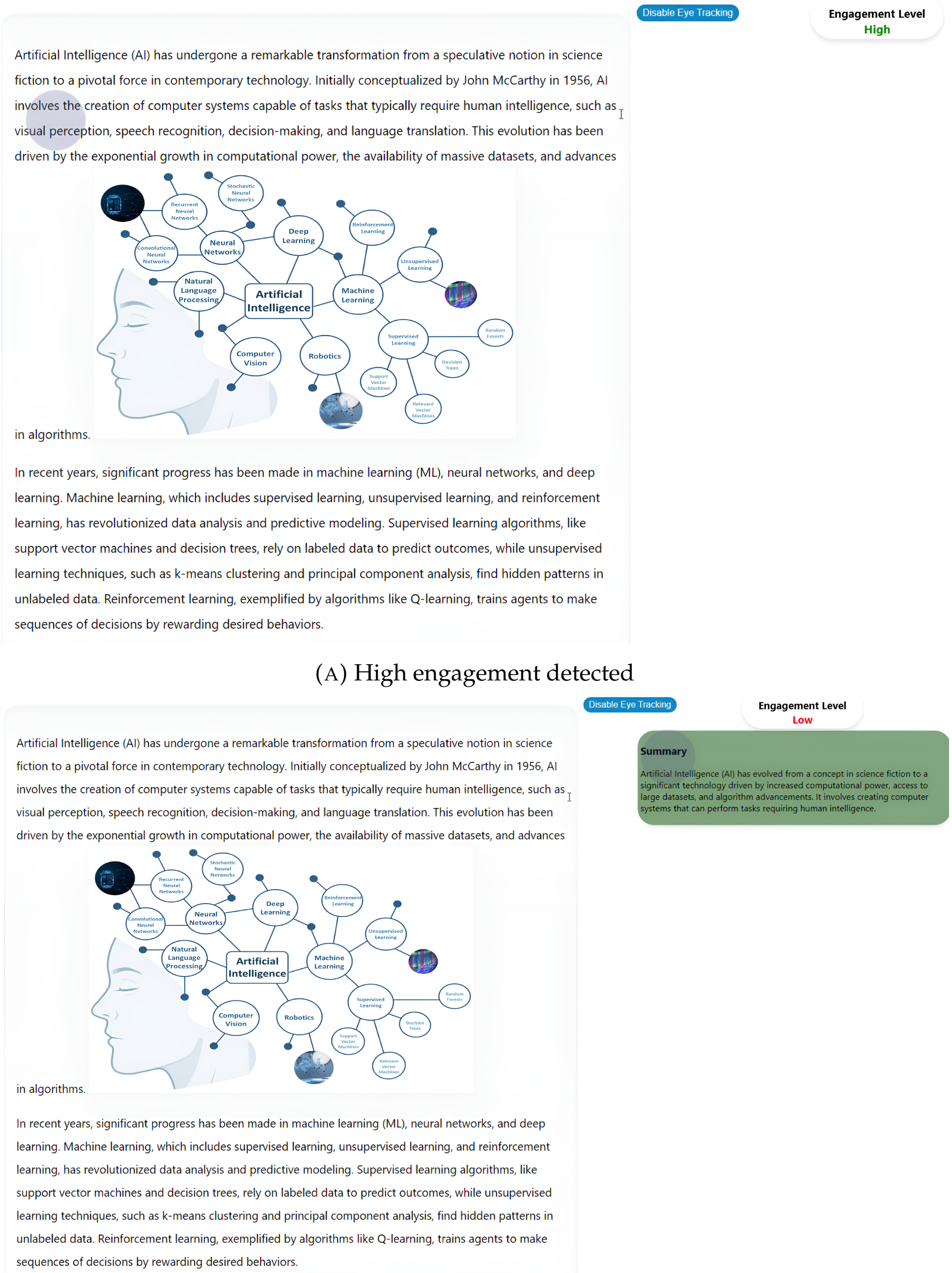
Where:

- S_{final} is the final selected summary.
- S_i is the i-th generated summary.
- $\{S_1, S_2, \ldots, S_n\}$ is the set of all generated summaries.
- $\text{Process}(S_i)$ is the function that includes the following steps:
  - Input Processing: The ChatGPT API processes the input paragraph with the prompt.
  - Summary Generation: The model generates multiple summaries based on the input.
  - Refinement: The generated summaries are processed to remove unwanted parts (e.g., URLs) and ensure they end with complete sentences.
  - Selection: The final summary is selected based on criteria such as coherence and relevance.

User Interface The user interface, built with React.js, ensures a smooth and intuitive experience for users. Participants start by reviewing the instructions and choosing their learning materials. After activating eye-tracking, fixation data and pupil diameter are transmitted to the engagement predictor every 5 seconds. The predicted level of engagement is displayed on a dashboard as either 'high' or 'low'. If the engagement level is indicated as 'low,' the system shows a summary of the content next to the paragraph being read. The layout of the system interface is illustrated in Figure 11. When high engagement is identified, the interface appears as shown in Figure 11(a). In contrast, when low engagement is detected, the system offers a summary of the content, as represented in Figure 11(b). This adaptive approach aims to re-engage learners and enhance comprehension by providing timely and relevant summaries [156].

Gaze Feature Metrics

In order to gain a deeper understanding of gaze behavior and attention trends, various features were derived from the raw gaze data. These features were chosen to



(A) High engagement detected

(B) Low engagement detected

FIGURE 11: The system interface with real-time engagement prediction dashboard and ChatGPT-based summary generation feature. (a) The system interface when high engagement is detected. (b) The system interface when low engagement is detected for a learning material.

highlight distinctions in gaze behavior and offer insights into both the experimental and control groups.

The extracted features included:

- **Fixation metrics**: mean fixation duration, standard deviation of fixation duration, fixation count, and mean pupil diameter, which provide information about the participant's ability to focus and maintain attention.

- **Saccade metrics**: mean saccade length, mean saccade angle, mean saccade velocity, and their standard deviations, which capture eye movement patterns, including speed and direction.
- **Blink metrics**: blink count and blink rate, which provide information about fatigue, distraction, or disengagement.

The analysis of these metrics aimed to detect patterns and trends that could guide the enhancement of adaptive learning systems. By examining these features, the study sought to understand how gaze behavior correlates with engagement and comprehension levels, and how the adaptive interventions influenced these metrics [156].

Model Training and Evaluation

To predict engagement and comprehension levels, two deep learning models were utilized: an InceptionTime network and a Transformer network. These models were chosen for their capability to analyze sequential data and identify both local and global patterns within time series data. The InceptionTime network adopts a multi-branch convolutional strategy to capture various temporal patterns. Each inception module consists of parallel branches consisting of convolutional layers with different kernel sizes, batch normalization, and activation functions. The modules are arranged in layers, followed by a global average pooling layer and a dropout layer to mitigate overfitting. The output layer at the end utilizes a sigmoid activation function for binary classification [156].

The Transformer network makes use of a multi-head attention mechanism to identify intricate temporal dependencies. It incorporates layers for multi-head attention, feed-forward processing, and residual connections, concluding with dense layers for classification. Dropout is used to prevent overfitting, and the final output layer employs a sigmoid activation function. To assess model performance, stratified group K-fold cross-validation with $K = 5$ was implemented. This technique ensures that each fold maintains the class proportions (low vs. high engagement/comprehension) while honoring the group structure (experimental vs. control). This method guarantees that the models are tested on representative samples and are applicable to both groups [156].

4.4.3 Results and Analysis

This subsection presents the findings of the study, focusing on the effectiveness of the adaptive learning system in enhancing learner engagement, comprehension, and self-confidence. The results are organized into three parts: gaze metrics analysis, survey response analysis, and predictive modeling analysis. The results and analysis are based on our published work [156].

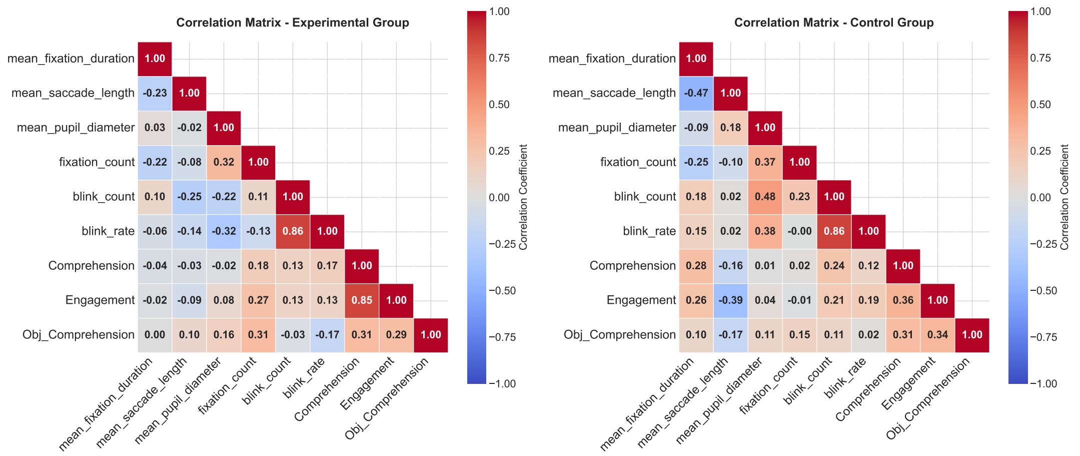


FIGURE 12: Correlation matrices of eye-tracking metrics, comprehension, engagement, and confidence for experimental and control groups.

Gaze Metrics Analysis

Eye-tracking metrics offer quantitative insights into how participants in both the experimental and control groups visually processed information. We extracted a variety of features to identify differences in gaze patterns between the groups. These features included average fixation duration, average saccade length, average saccade angle, average saccade velocity, standard deviation of fixation duration, standard deviation of saccade length, standard deviation of saccade angle, standard deviation of saccade velocity, average pupil diameter, number of fixations, number of saccades, number of blinks, and blink rate. These metrics provide a detailed view of participants' visual behavior during the experiment [156].

Figure 12 presents correlation matrices that illustrate the relationships between eye-tracking metrics and cognitive outcomes (comprehension and engagement) for both the control and experimental groups. In the experimental group, the correlations between gaze metrics and engagement or subjective comprehension are weaker. This suggests that the adaptive interventions change the typical relationship between eye movements and cognitive responses. For instance, while longer fixation durations correlate with better comprehension and engagement in the control group, the interventions in the experimental group seem to promote more strategic reading patterns. Additionally, saccade length shows a positive correlation with objective comprehension in the experimental group, indicating that participants who used the generated summaries achieved higher comprehension levels compared to those in the control group [156].

Figure 13 presents a detailed analysis of gaze metrics across different comprehension and engagement levels. The control group exhibits increased mean fixation duration alongside higher engagement and comprehension, with saccade length decreasing indicative of typical reading behaviors. Conversely, the experimental group displays more consistent metrics, such as relatively stable fixation durations and gradual saccade length trends. This suggests the intervention helps stabilize these

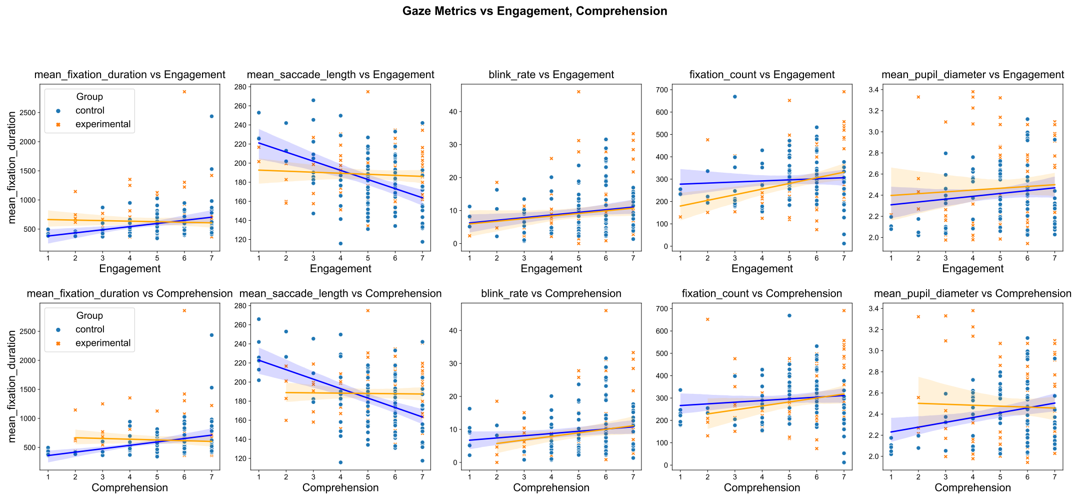


FIGURE 13: Comparison of eye-tracking metrics between control and experimental groups across engagement and comprehension levels (1-7)

metrics, maintaining participant focus on relevant content. The experimental group shows elevated fixation counts and larger pupil diameters when comprehension and engagement are lower, pointing to greater cognitive effort and arousal resulting from the adaptive interventions [156].

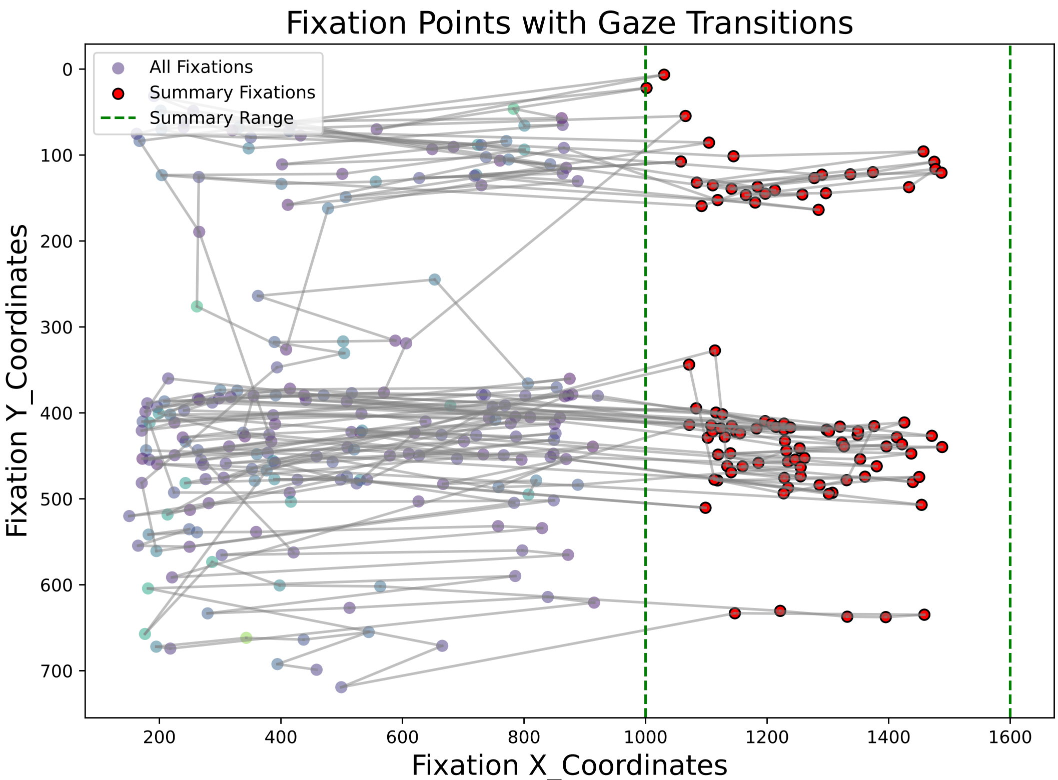


FIGURE 14: The transition of fixations between the actual text and generated summary for a participant in the experimental group detected with low engagement.

Figure 14 illustrates how a participant in the experimental group shifted their focus between the original text and the summaries generated for them. This visualization highlights the effectiveness of the summaries in capturing the participant's attention and encouraging a return to the main content with better focus [156].

Survey Response Analysis

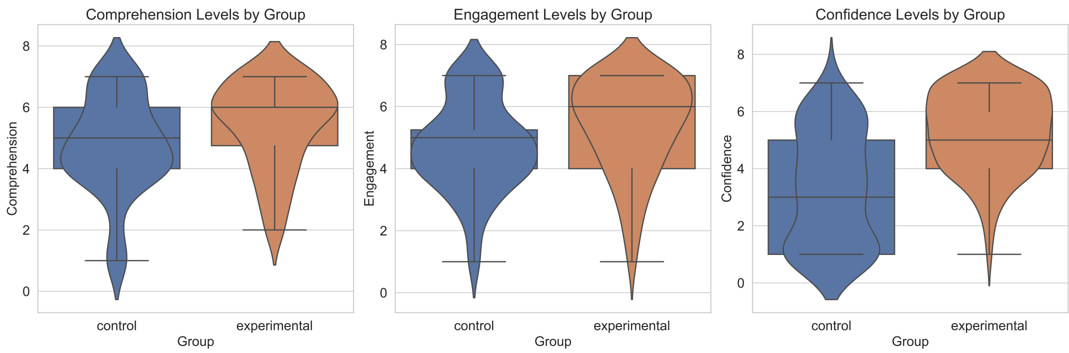


FIGURE 15: The distribution of subjective comprehension, engagement, and confidence levels by group (control vs. experimental)

After the reading sessions, participants completed surveys to rate their comprehension, engagement, and confidence. Figure 15 presents the distribution of these ratings for both the control and experimental groups. Participants using the adaptive learning system reported greater engagement and confidence, with some improvement in perceived comprehension. T-tests confirmed statistically significant differences between the groups [156]:

- Engagement: $t = 3.12, p = 0.0021$
- Confidence: $t = 8.22, p < 0.0001$
- Comprehension: $t = 3.17, p = 0.0018$

Figure 16 illustrates the mean values and variability in subjective ratings. It reveals that participants using the adaptive system reported greater engagement and confidence, with a slight increase in comprehension. The range of responses indicates that while the system generally worked well, individual experiences varied [156].

To rigorously test our hypothesis that the adaptive learning system enhances objective comprehension and to mitigate potential biases from subjective comprehension ratings, we administered an objective comprehension quiz for each document. Each quiz consisted of four questions specific to the document's content. The results from these objective assessments indicated notable performance differences between the control and experimental groups, largely supporting our hypothesis. Figure 17a illustrates the distribution of total correct answers for both groups, while Figure 17b presents the overall distribution of correct answers across the groups. A greater proportion of participants in the experimental group achieved perfect scores

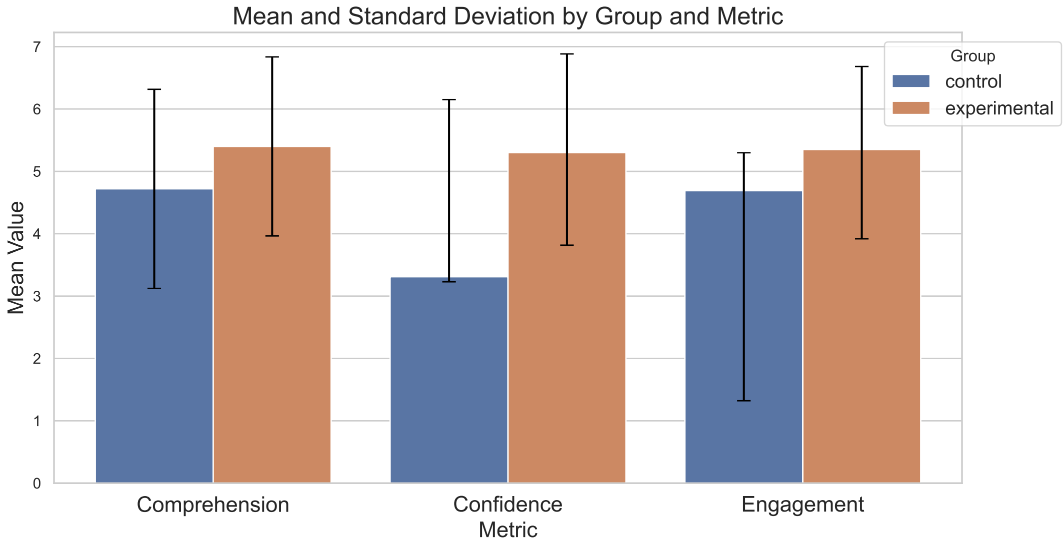


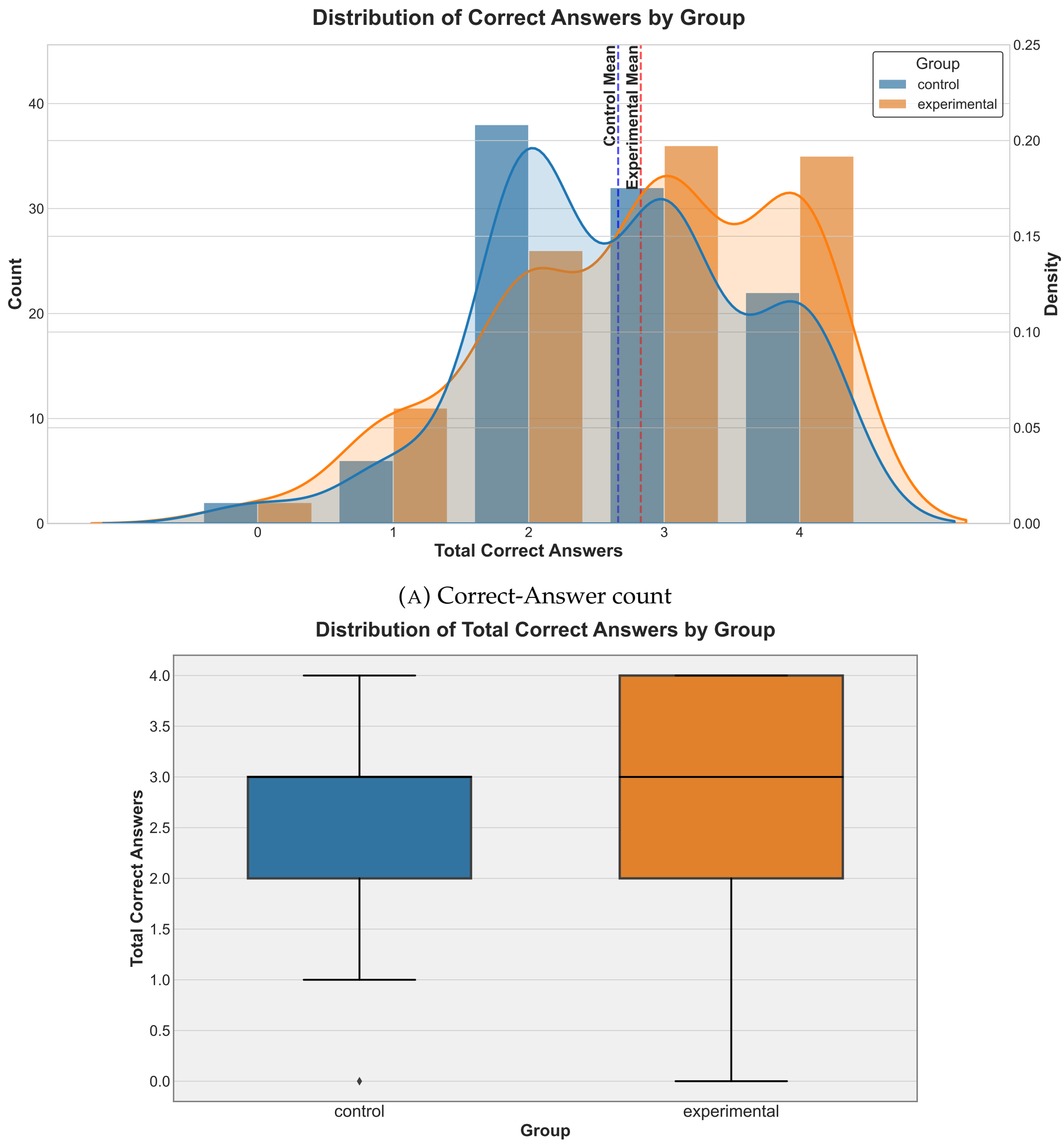
FIGURE 16: Mean and standard deviation of Comprehension, Confidence, and Engagement by group.

(4/4), suggesting that the adaptive interventions effectively boosted comprehension for a substantial segment of this group. Although the experimental group demonstrated an overall improvement, some participants within this group still had lower scores (1/4 or 2/4). This variability indicates that the adaptive interventions, particularly the generated summaries, may not have had a consistent impact across all participants. Several factors could explain this inconsistent effectiveness, including individual learning preferences, prior familiarity with the subject matter, or varying degrees of engagement with the adaptive features, all of which could influence how participants benefited from the summaries. Despite these variations, the higher incidence of perfect scores in the experimental group supports our hypothesis that the adaptive learning system generally enhances objective comprehension. These findings not only validate the potential of our approach but also highlight areas for future refinement to ensure more consistent benefits across diverse learners [156].

Predictive Modeling Analysis

To predict engagement and comprehension ratings, two deep learning models were employed: InceptionTime and a Transformer network. These models were trained using raw gaze data, including x and y coordinates, as well as pupil diameter measurements. The task was structured as a binary classification problem, where ratings from 1 to 4 were categorized as 'low' and ratings from 5 to 7 were classified as 'high'. To ensure a balanced representation of both groups and classes in each fold, we utilized a 5-fold StratifiedGroupKFold cross-validation strategy [156].

Table 4.5 presents a summary of the evaluation metrics for both models. The Transformer model consistently showed better performance than the InceptionTime model, with higher accuracy and F1-scores in predicting both engagement and comprehension. Specifically, the Transformer model achieved 68.15% accuracy and a



(A) Correct-Answer count

(B) Objective Comprehension Distribution

FIGURE 17: Distribution of total correct answers (out of 4 questions) for objective comprehension across control and experimental groups.

65.82 F1-score for engagement prediction, while the InceptionTime model reached 62.71% accuracy and a 60.45 F1-score. For comprehension prediction, the Transformer model attained 69.60% accuracy and a 64.32 F1-score, whereas the InceptionTime model achieved 60.32% accuracy and a 59.18 F1-score. The engagement accuracy for both models over the 5-fold cross-validation is illustrated in Figure 18. Throughout every fold, the Transformer model consistently outperformed the performance of InceptionTime. This highlights the Transformer model's capability to understand long-range dependencies in sequential data, making it especially effective for examining gaze patterns over time [156].

TABLE 4.5: Summary of evaluation metrics for Engagement and Comprehension

Models	Engagement		Comprehension	
	Accuracy	**F1-score**	**Accuracy**	**F1-score**
InceptionTime	62.71	60.45	60.32	59.18
Transformers	**68.15**	**65.82**	**69.60**	**64.32**

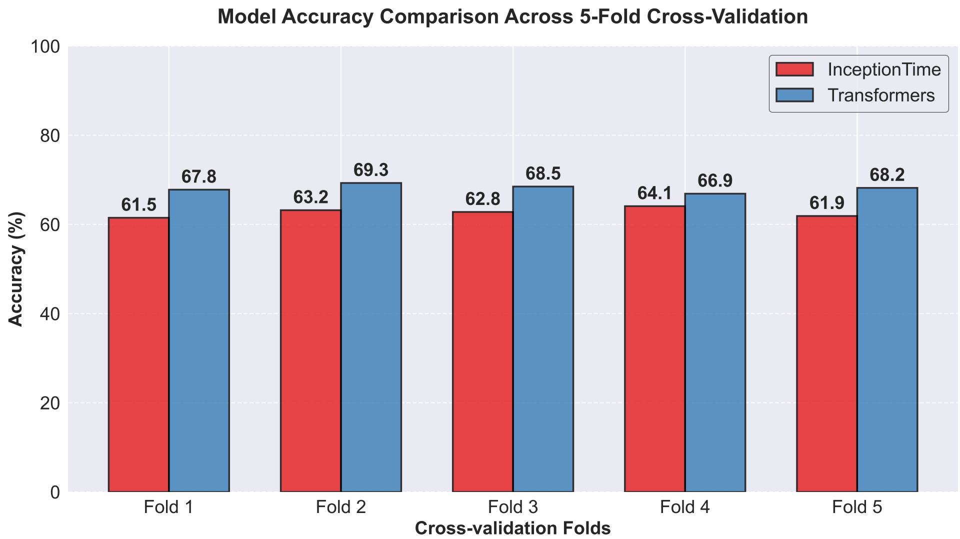


FIGURE 18: Model Accuracy Comparison Across 5-Fold Cross-Validation

4.4.4 Discussion and Implications

This section delves into the study's significant contributions, results, and insights, combining both the theoretical implications and practical uses of the research. It emphasizes the novel real-time adaptive system's influence on learning results and its relevance to current discussions in the field. Furthermore, the study's challenges and limitations are critically assessed. The discussion and implications are based on our published work [156].

Gaze-Based Real-Time Engagement Prediction

The user interface was designed to capture participants' gaze data in real time and dynamically display predicted engagement levels. An interactive dashboard visually represented these engagement levels, which were predicted every 5 seconds using pre-processed gaze data, including fixation and pupil diameter metrics. To evaluate the system's effectiveness, the number of summaries displayed to participants in the experimental group was recorded after they read each document. This allowed for a direct correlation between the number of summaries and user-reported engagement ratings, as well as an analysis of how post-survey responses aligned with the number of summaries generated. Additionally, participants in the experimental group rated the system's ability to predict gaze behavior in the post-survey, providing insights into their perceived engagement with the documents [156].

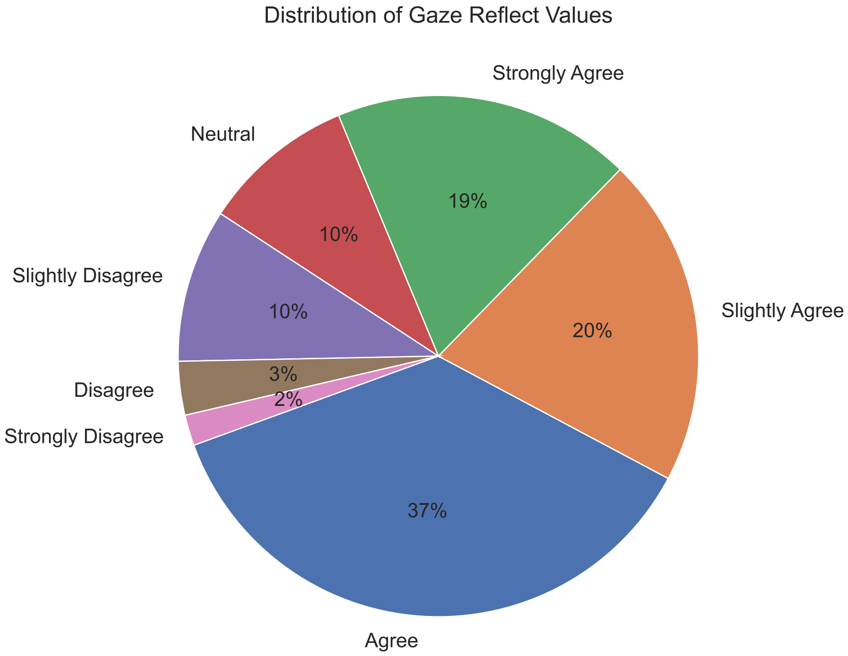


FIGURE 19: Distribution of ratings provided by participants reflecting the system's effectiveness in predicting gaze behavior for engagement prediction.

Figure 19 illustrates participant feedback on the system's ability to predict gaze behavior for engagement, using a scale of 1 to 7 (where 1 = 'strongly disagree' and 7 = 'strongly agree'). The results indicate that a large portion of participants—approximately 37% (rating of 6, indicating agreement) believed the system effectively captured their gaze patterns and accurately predicted their engagement. An additional 19% strongly agreed, and 20% slightly agreed with the system's predictive capabilities. In contrast, only a small minority expressed disagreement, with 2% strongly disagreeing and 3% disagreeing. Overall, these findings suggest that most participants found the system to be effective in predicting their real-time engagement levels based on gaze behavior [156].

The number of summaries used by the experimental group for each document was tracked and analyzed in relation to their self-reported engagement levels. Figure 20 reveals that the most summaries were, on average, accessed when participants reported the lowest engagement (level 1). This suggests that the system effectively provided support when participants were struggling. As engagement increased from levels 1 to 7, the average number of summaries used decreased. This indicates that more engaged participants relied less on the extra support, likely because they understood the material better on their own. The fewest summaries were used at the highest engagement levels (6 and 7), suggesting that participants at these levels were confident in their comprehension and didn't need additional assistance.

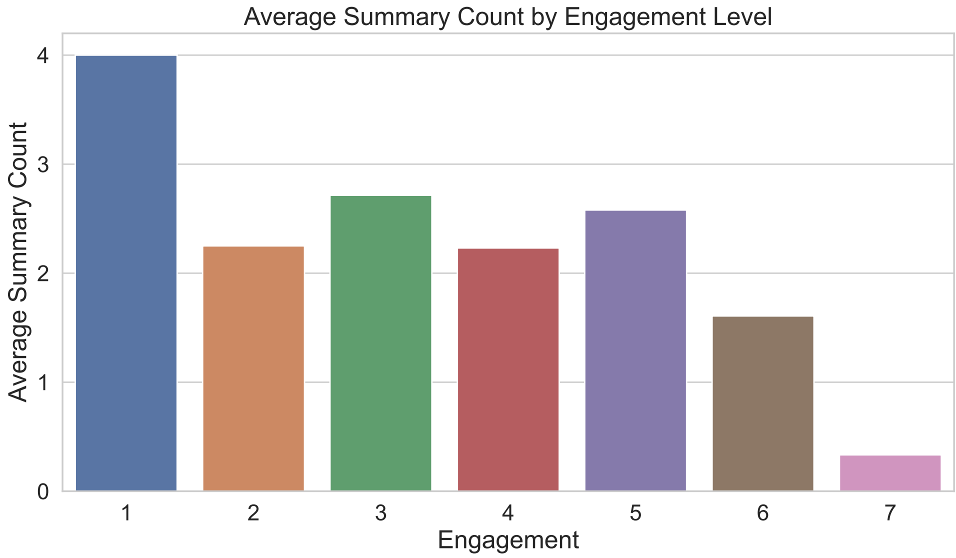


FIGURE 20: The average summary count corresponding to different levels of engagement as rated by participants.

This inverse relationship highlights the system's adaptive nature, providing support based on real-time engagement and demonstrating the importance of monitoring engagement to effectively tailor educational interventions [156].

The system's accuracy in predicting engagement from gaze data reached 68.15%, with an F1-score of 65.82. This performance is on par with other current methods in the field. For instance, Chen et al. used a multi-modal deep neural network (MDNN) that combined gaze direction and facial expressions to predict student engagement in collaborative settings, achieving high accuracy [30]. Sharma et al. developed an engagement index based on eye movements, head movements, and facial emotional cues, which effectively classified students into varying engagement levels [169]. While Gupta et al. attained higher accuracy rates using deep learning architectures like ResNet-50 for facial emotion analysis, these techniques often demand intricate setups and are constrained by strict front-facing requirements [63]. Our system, however, depends solely on gaze data, providing comparable accuracy without being affected by head position or orientation. This makes it especially useful in authentic learning environments [156].

In contrast to studies that focus solely on model accuracy, our system employs a dual validation strategy. This includes technical validation via model performance metrics and user-centered validation, where a substantial majority of participants (76%) agreed with the system's engagement predictions. Our system offers real-time visualization through an interactive dashboard, providing continuous engagement monitoring at 5-second intervals. This contrasts with prior research that used longer intervals or post-hoc analysis. This immediate feedback, coupled with automated interventions, establishes a learning loop where users can instantly confirm and benefit from the system's assessments. Despite achieving slightly lower technical accuracy compared to multimodal approaches, the high rate of user agreement suggests that real-time feedback and immediate interventions may offer more practical value in learning environments than the incremental gains in prediction accuracy achieved by more complex systems [156].

Impact of the System on Learning Outcomes

The system was developed to provide summaries only when it detected low levels of real-time engagement, offering tailored assistance to learners. This method aimed to reengage participants during periods of disengagement and enhance both engagement and comprehension through customized interventions. While earlier research on adaptive e-learning environments [50] showed significant advancements in student engagement through personalization based on learning styles, our system adopts a different method by applying real-time, engagement-based interventions via automated summary generation. The summaries, created using a ChatGPT-based API, delivered brief, contextually relevant information to assist learners in understanding the material better. This approach is particularly beneficial for learners who may struggle with maintaining focus or comprehending complex content, as it provides immediate support tailored to their current engagement levels [156].

Feedback collected through surveys indicated that the experimental group, which utilized the adaptive learning system, experienced greater engagement and confidence compared to the control group, along with some improvement in subjective comprehension. Statistical analysis validated these observations, revealing significantly greater engagement ($t = 3.12, p = 0.0021$) and confidence ($t = 8.22, p < 0.0001$) levels in the experimental group, as well as a significant enhancement in comprehension ($t = 3.17, p = 0.0018$). These results are consistent with the findings of Liu et al. [111], which identified positive relationships between learning achievement, engagement, and learning attitudes. However, the variability in responses implies that while the adaptive system proved advantageous for numerous participants, its effectiveness varied among learners. Some individuals found the summaries especially beneficial, while others did not perceive the same level of advantage. This variability underscores the necessity of customizing educational interventions to accommodate a range of learner needs and preferences, ensuring that adaptive systems can effectively support diverse learning styles and requirements [156].

An analysis of gaze metrics was conducted to quantitatively assess the gaze behavior of learners in both the experimental and control groups. Participants in the experimental group frequently utilized the summaries as interventions when their engagement levels were low, suggesting that these summaries helped shift their focus. The pupil diameter and fixation durations in the experimental group were substantially higher during periods of lower engagement and comprehension than those in the control group. This increase suggests that participants in the experimental group experienced elevated cognitive processing when interacting with the generated summaries. The larger pupil diameter often correlates with increased cognitive load and arousal, indicating active participation with the material to enhance understanding. When engagement levels drop, the introduction of summaries seems to enable participants to concentrate more on the content, facilitating their ability to extract relevant information effectively. The prolonged fixation duration indicates a deeper engagement with the summaries, as participants devote more time to process the information for better comprehension. This cognitive effort, driven by the summaries, functions as a beneficial intervention that aids in reengaging learners and reinforces their grasp of the material [156].

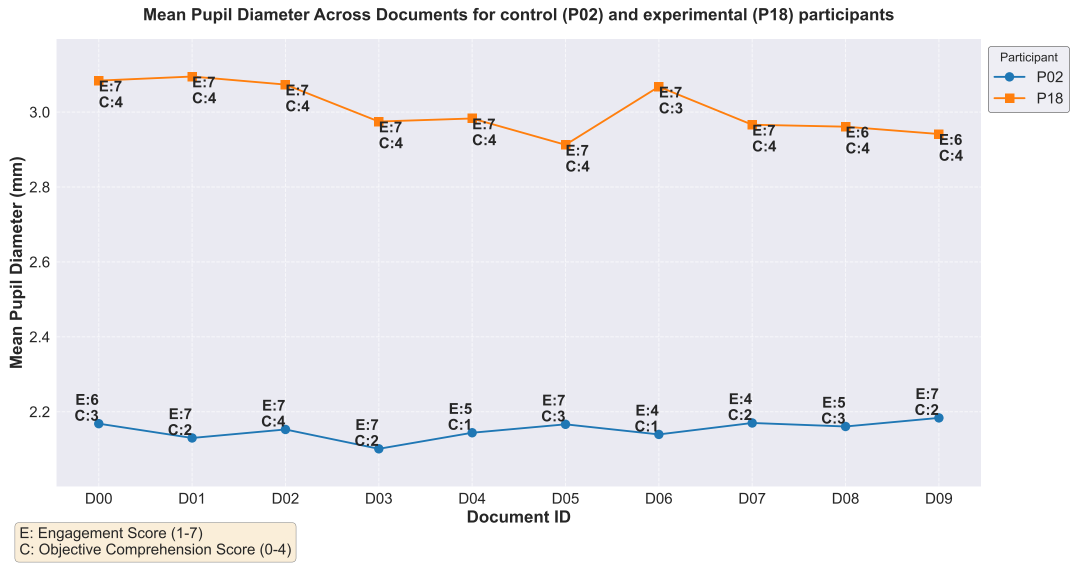


FIGURE 21: Variation in pupil diameter across control (P02) and experimental (P18) participants over the learning material, with reported engagement and computed objective comprehension scores.

Figure 21 shows the variation in mean pupil diameter for participants in the control (*P02*) and experimental (*P18*) groups. The experimental participant (*P18*) generally showed a larger average pupil diameter, suggesting increased cognitive effort or sustained focus, which is consistent with their higher engagement and comprehension scores. Conversely, the control participant (*P02*) displayed smaller pupil diameters and more fluctuating engagement and comprehension scores, indicating variable involvement. These differences highlight the potential advantages of the adaptive interventions received by the experimental group, which may have contributed to more consistent cognitive engagement and better learning outcomes [156].

Although statistical significance was not achieved in the gaze metric comparisons between the experimental and control groups, the observed patterns offer valuable insights into participant behavior. Gaze metrics, such as fixation duration and pupil diameter, suggested that participants in the experimental group interacted with the generated summaries, potentially indicating increased cognitive processing, even though the differences were not statistically significant. This lack of statistical significance may reflect the variability in individual responses to the summaries, suggesting that while the adaptive learning system's impact on gaze behavior may not have been uniform across all participants, it still shows promise as a tool for enhancing engagement and comprehension. Overall, these findings highlight the potential benefits of adaptive learning interventions in promoting deeper cognitive engagement and improving learning outcomes, even in challenging contexts [156].

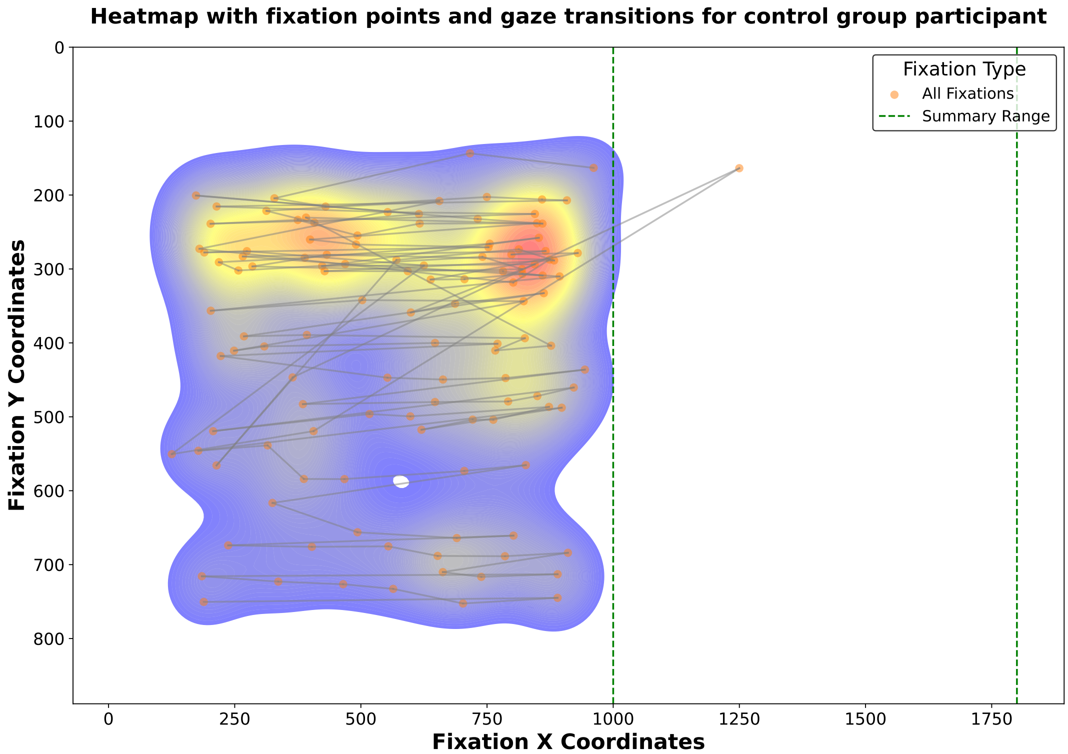
Visual effort and cognitive outcomes

To understand the connection between cognitive processing, visual effort, and adaptive learning, it's crucial to pinpoint key metrics that reflect cognitive and visual engagement. Fixations, saccade patterns, and gaze transitions each provide unique insights into a reader's cognitive processes and visual strategies. Research has demonstrated that eye-tracking metrics like fixations, saccades, and pupil diameter are reliable indicators of visual attention and cognitive workload [176] [156].

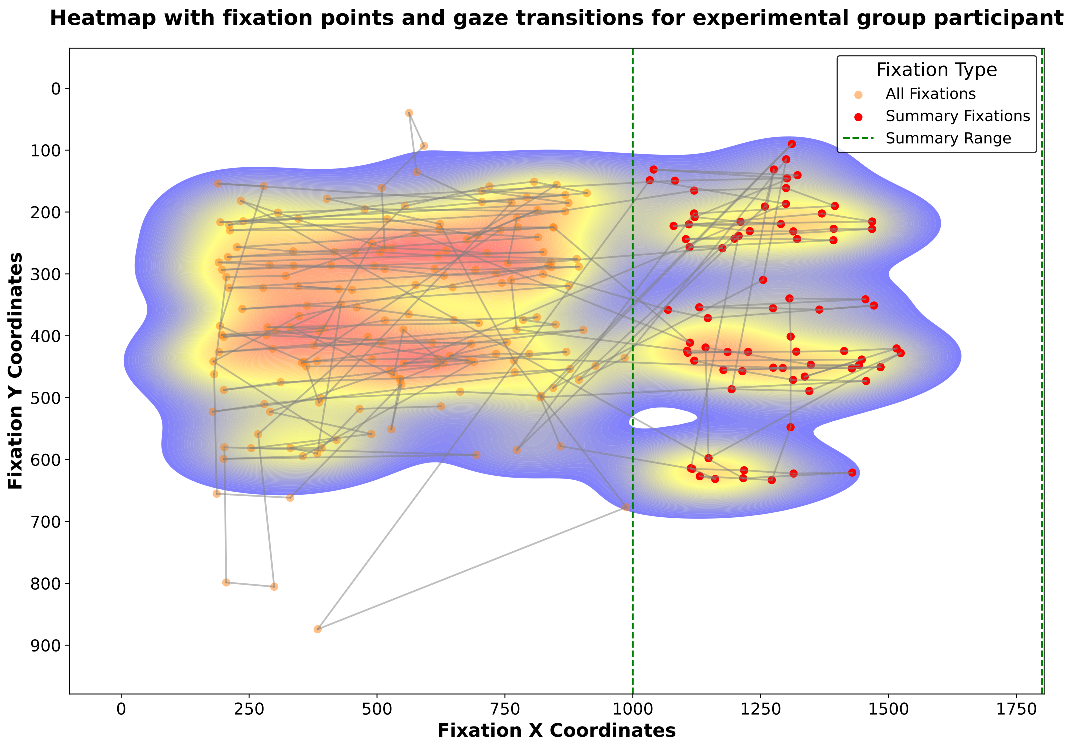
Our study uses control and experimental groups to compare how people read and solve problems. Similar designs exist, but we go further. Prior work used eye-tracking to understand problem-solving [76] or cognitive effort during tests [174]. Unlike these post-hoc analyses, we predict engagement in real-time and adapt the learning experience using gaze patterns. The control group reads naturally, establishing baseline behaviors. The experimental group receives personalized help to optimize their learning [156].

Since the experimental group alone received adaptive interventions, we expect their interactions with the learning materials to differ. Consequently, we anticipate variations in gaze patterns, fixation times, and visual exploration compared to the control group. The control group's eye movements should reflect conventional reading, while the experimental group may show more focused visual strategies due to the adaptive interventions [156].

Figure 22 presents heatmaps showing fixation points and gaze transitions for a participant from the control group and one from the experimental group as they read the same material, both reporting low engagement afterward. The experimental group's heatmap shows fixations on the summaries (indicated by red dots within the green dashed lines), a feature absent in the control group's data. This difference directly reflects the adaptive intervention. For the experimental participant, this adaptive feature provides targeted support by offering a concise overview of key information when engagement drops, allowing for a personalized learning pace and enabling participants to quickly catch up for better content understanding and engagement recovery. These summary fixations suggest that participants are using this additional resource for potentially more efficient comprehension when their engagement level decreases. Conversely, the control group's gaze patterns likely reflect a more uniform but potentially less responsive interaction with the material. This contrast highlights the potential of adaptive interventions to deliver timely, targeted support based on real-time engagement levels [156].



(A) Control Group Participant



(B) Experimental Group Participant

FIGURE 22: Heatmaps of fixation points and gaze transitions for control and experimental group participants reading the same material with low reported engagement levels.

The patterns of visual effort and cognitive processing diverge significantly between the control and experimental groups. The participant in the control group

(Figure 22a) displays more focused fixations and briefer gaze transitions, suggesting a potentially reduced visual effort but also a more conventional, linear reading style. This pattern may indicate a consistent level of cognitive demand throughout the task. Conversely, the participant in the experimental group (Figure 22b) exhibits a broader distribution of fixations and longer gaze transitions resulting from the interventions, indicating potentially greater visual effort as they explore more of the content. This heightened visual exploration likely corresponds to more varied cognitive strategies and fluctuating levels of cognitive demand. The adaptive interventions provided to the experimental group seem to encourage a more dynamic interaction with the content, potentially leading to deeper cognitive engagement. While the experimental group may expend more visual effort, this increased activity could signify more thorough information processing and potentially more effective learning. The relationship between visual effort and cognitive processing in this context suggests that the adaptive interventions may be fostering a more active and comprehensive approach to learning, albeit with increased visual demands [156].

Potential Impact on Educational Practice

This study's results offer significant implications for educational settings, especially in online and self-directed learning. By providing customized, real-time support, the system caters to diverse student requirements, offering extra help or advanced material as needed. Educators can use the system to identify struggling students, enabling timely adjustments to teaching methods. Engagement and gaze data can inform curriculum design and support teacher development. The system's monitoring of engagement and comprehension offers a detailed view of learning, complementing traditional assessments. By adapting to varied learning needs, the system supports students with different attention levels, promoting inclusive education [156].

By merging eye-tracking with AI-generated summaries, this system provides adaptable learning assistance that can be scaled and complements standard evaluations. This method can promote deeper cognitive involvement and enhance learning results, especially for students with attention difficulties or those needing customized support. The system's monitoring of engagement and comprehension offers a detailed view of learning, supplementing traditional assessments. By adapting to varied learning requirements, the system aids students with differing attention spans, fostering inclusive education. This approach has the potential to transform educational practices by making learning more responsive and personalized, ultimately leading to improved student outcomes and satisfaction [156].

4.4.5 Limitations and Future Work

The adaptive learning system utilizing eye-tracking technology provides important insights, but several limitations were recognized, each opening opportunities for future research. The limitations and potential future directions are discussed below. The limitations and future work are based on our published work [156].

Sample Size and Representation: A significant limitation of this research is the limited sample size (N=22). This limitation may impact the ability to generalize our findings to larger populations. The challenge is to obtain a sample that is sufficiently large and diverse to accurately reflect various learning styles, backgrounds, and cognitive abilities [156].

Content and Real-Time Engagement Detection: The specific tasks or content employed in this study may not be equally engaging or challenging for all participants, which could have a bias in the findings. The task of producing uniformly appropriate content for different skill levels and interests presents a notable challenge. The precision and responsiveness of the system in detecting decreased engagement in real-time may be constrained by relying solely on gaze data; integrating multimodal data, including facial expressions and physiological indicators, could enhance accuracy. Creating algorithms capable of reliably interpreting eye movements as signs of engagement across varying individuals remains a complicated task [156].

Adaptive Intervention Strategy and Long-term Effects: The use of summaries as the only method of intervention for low engagement may have limitations. Designing interventions that are effective for all learning styles and preferences is a challenging task. Furthermore, this study primarily focuses on immediate outcomes without taking into account long-term learning results. The difficulty lies in creating longitudinal studies that can evaluate the impact of these interventions on knowledge retention and application over time [156].

LLM-based Summaries: The quality and relevance of the AI-generated summaries may vary, potentially affecting their effectiveness as learning aids. Ensuring that these summaries are accurate, concise, and contextually appropriate is a significant challenge. The reliance on AI-generated content also raises concerns about the potential for misinformation or oversimplification of complex topics, which could hinder rather than help learning [156].

Individual Differences and Participant Responses: Individual differences in learning preferences, prior knowledge, and cognitive abilities may influence how participants respond to the adaptive interventions. The challenge is to design a system that can effectively accommodate these differences and provide personalized support that meets the unique needs of each learner [156].

Future work should address several key areas to enhance the adaptive learning system. Increasing the sample size and ensuring a more diverse participant pool will improve the generalizability of the findings. Future studies should also explore a broader range of content types and difficulty levels to better understand how different materials impact engagement and comprehension. Integrating multimodal data sources, such as facial expressions and physiological signals, could improve the accuracy of real-time engagement detection. Developing more sophisticated algorithms that can interpret eye movements in conjunction with other indicators of engagement will be crucial [156].

Furthermore, future studies could explore a wider variety of adaptive intervention strategies beyond summaries, such as interactive quizzes, visual aids, or personalized feedback. Longitudinal studies are needed to assess the long-term effects of these interventions on learning retention and application. Ensuring the quality and relevance of AI-generated content will be essential, potentially involving human oversight or validation processes. Finally, future research should focus on understanding and accommodating individual differences in learning styles and cognitive abilities to create a more personalized and effective adaptive learning experience [156].

4.4.6 Conclusion

This study demonstrates the potential of an eye-tracking-based adaptive learning system to enhance student engagement and comprehension through real-time interventions. By leveraging gaze data to predict engagement levels, the system provides personalized support in the form of AI-generated summaries when engagement drops. The experimental group, which utilized this adaptive system, reported higher engagement and confidence levels, along with improved comprehension compared to the control group. Gaze metrics indicated that participants in the experimental group exhibited increased cognitive effort when interacting with the summaries, suggesting that these interventions effectively reengaged learners and facilitated deeper information processing. These findings highlight the promise of integrating eye-tracking technology with adaptive learning strategies to create more responsive and personalized educational experiences. However, the study also acknowledges limitations such as sample size and the need for diverse content, suggesting avenues for future research to further refine and validate this approach [156].

Chapter 5
Conclusion

This chapter provides a conclusion to the thesis by summarizing its key aspects and outlining future directions. Section 5.1 offers a brief overview of the thesis, highlighting its main findings. Section 5.2 details the contributions made to the field through this research. Section 5.3 discusses the limitations of the research and proposes potential directions for future research to build upon this work.

5.1 Summary

This thesis explored the enhancement of digital learning through three interconnected phases: detection, intervention, and adaptation. These phases were guided by three key research questions, each addressing a critical aspect of personalized e-learning. The following sections summarize how each research question was answered through the research conducted in this thesis.

5.1.1 How can sensor data and deep learning be utilized to effectively detect learners' affective and cognitive states?

The first research question was addressed by developing a comprehensive framework for detecting learners' affective and cognitive states using multimodal sensor data and deep learning models. Chapter 3 presented the development of advanced deep learning architectures designed to process data from multiple sensors, including eye-trackers, physiological signals, and computer interaction patterns to detect learners' affective and cognitive states.

Section 3.1 investigated stress detection using sensor-based physiological data from the WESAD and SWELL-KW datasets (public), focusing on evaluating and comparing traditional machine learning approaches and deep learning approach. For the WESAD dataset, a Fully Convolutional Network (FCN) outperformed traditional models, achieving 93.54% accuracy and 91.39% F1-score for binary stress classification using chest-worn physiological signals. This demonstrated the strength of deep learning in automatically learning complex patterns from high-dimensional data without manual feature extraction. For the SWELL-KW dataset, the analysis revealed that combining multiple modalities improved performance, achieving

91.24% accuracy for three-class classification in user-dependent scenarios. A significant performance gap (approximately 45%) between user-dependent and user-independent approaches underscored the highly individualized nature of stress responses. These findings demonstrate the effectiveness of deep learning with rich sensor data for detecting affective and cognitive states while also emphasizing the challenges of user variability and the need for personalized approaches.

Section 3.2 demonstrated the effectiveness of deep learning models, ResNet and LSTM, in detecting stress levels (easy, medium, hard) in a controlled environment using physiological signals from an Empatica E4 wristband. ResNet achieved 95.05% accuracy with KFold, while LSTM reached 91.17% accuracy with 20% fine-tuning with LOPO. User-specific fine-tuning consistently improved performance, highlighting its importance for personalized stress detection and the potential of deep learning in this domain.

Section 3.3 investigated interest detection while reading newspaper articles using a multimodal approach combining gaze data and physiological signals. A CNN-LSTM model achieved 82.3% accuracy for binary classification, outperforming manual feature extraction methods. A person-dependent approach further improved accuracy, highlighting the importance of personalized predictions. These findings demonstrate the potential of deep learning and multimodal data for understanding interest, with applications in personalized learning and educational settings.

Section 3.4 addressed the detection of engagement and emotion levels during reading tasks using deep learning models and multimodal data from an eye-tracker and E4 wristband. The study demonstrated that deep learning models, such as Transformers and ResNet, effectively predict engagement, valence, and arousal, with Transformers achieving 80.38% accuracy for engagement in user-independent settings and ResNet excelling in user-dependent settings with 93.56% accuracy. The findings revealed strong correlations between engagement and emotions, highlighting the unique affective responses evoked by different texts.

Section 3.5 explored how sensor data, specifically gaze features and fixation heatmaps, can be used to assess document readability and understand cognitive states during reading. The study demonstrated that features such as saccade length, regression ratio, and fixation duration provide valuable insights into reading behavior and document difficulty. By integrating gaze data with text features, the research highlighted the potential of sensor-based approaches to detect cognitive states like readability and engagement.

5.1.2 How to design and deliver personalized interventions that effectively address learners' mental states and improve their learning experience?

The second research question was addressed by developing personalized intervention strategies tailored to learners' cognitive and emotional states. Chapter 4 explored methods for designing and delivering interventions that leverage real-time predictions of mental states, such as stress, engagement, and interest, to enhance learning experiences. These strategies included controlled distraction, cognitive training, emotional engagement, and adaptive systems, ensuring that interventions dynamically respond to individual needs and improve user interaction and performance.

Section 4.1 introduced a novel approach to detect and model distraction during reading tasks using eye-tracking data. A counting distractor task objectively captured distraction, revealing its impact on reducing interest and increasing perceived difficulty. Machine learning models predicted distraction and interest levels, providing insights for designing interventions to monitor and mitigate distraction, enhancing focus and engagement in real-world settings.

Section 4.2 investigated the impact of notifications and pop-ups on reading comprehension and frustration, demonstrating that frequent interruptions significantly hinder comprehension and increase frustration. Cognitive control training emerged as a key intervention, significantly improving comprehension and reducing distraction levels. These findings highlight the potential of cognitive training and thoughtful interface design as personalized interventions to enhance focus and reading performance in distraction-heavy digital environments.

Section 4.3 introduced a personalized feedback system that enhances user engagement and emotional states through an interactive dashboard and tailored interventions. The system integrates data collection, machine learning predictions, and user feedback, using a user-dependent approach for high personalization. The dashboard provides real-time feedback through features like an engagement gauge, emotion emoji, and motivational messages, actively improving focus and emotional well-being. This system has applications in e-learning, workplace productivity, and mental health, offering a foundation for adaptive technologies that prioritize user engagement and emotional health.

5.1.3 How to adapt learning materials to learners' real-time mental states?

The third research question was addressed by developing an adaptive learning system capable of dynamically adjusting learning materials based on learners' real-time mental states. This approach ensures that learning experiences are personalized, responsive to individual needs, and optimized to improve overall engagement and performance.

Section 4.4 demonstrated the effectiveness of integrating real-time gaze tracking with AI-driven adaptive learning interventions to enhance engagement and learning outcomes. Adaptive interventions, triggered by detected low engagement, significantly improved both engagement and comprehension levels. Deep learning models, including a Transformer network, successfully predicted post-reading engagement and comprehension from gaze patterns, with the experimental group achieving higher comprehension scores compared to those without adaptive interventions. This research highlights the feasibility of using gaze data for real-time engagement prediction and provides a framework for integrating eye-tracking with AI-driven adaptive learning systems, advancing personalized and responsive educational technologies.

5.2 Contributions to the Field

This research makes significant contributions to the field of e-learning by advancing the understanding, modeling, and application of adaptive learning systems tailored to individual learners' needs. The integration of deep learning approaches and sensor fusion techniques has enhanced the accuracy of predicting learners' affective and cognitive states, which are critical for creating personalized and effective learning experiences. By refining the computation of these states, the research achieves robust results in both generalized and personalized validation approaches, demonstrating the potential for scalable and adaptable solutions in diverse learning environments.

The development of personalized intervention techniques and adaptive learning systems represents an impactful step in the e-learning ecosystem. These contributions enable a smooth interaction between the learner and the environment, providing a more engaging and supportive learning experience. By addressing individual differences in learning behavior, the research motivates learners to overcome challenges, reduce stress, and maintain motivation, ultimately improving learning outcomes.

The visual feedback system, developed based on sensor data and affective/cognitive states, provides a novel application to enhance user engagement and self-awareness in the learning process. This application incorporates features such as engagement and stress indicators, alert messages, emotion-based visual elements, and motivational messages, offering learners a reflective tool to better understand their own learning behaviors and emotional states. This system serves as a valuable resource for learners to analyze their interactions and emotional responses during the learning process. Additionally, it provides educators and system designers with insights into learner behavior, enabling them to design more effective and personalized learning strategies.

This work contributes to the growing body of knowledge in adaptive e-learning by bridging the gap between advanced computational techniques and practical applications. By leveraging real-time gaze data, the system detects engagement levels using a pre-trained model, enabling AI-driven adaptations to the content in real time. This real-time adaptability ensures that learners remain engaged and receive personalized support as they progress through the learning process. The research lays the foundation for future advancements in personalized learning by offering a framework for creating systems that are both effective and learner-centric. These contributions have the potential to shape the future of e-learning, making it more inclusive, efficient, and enjoyable for learners across diverse contexts. By integrating real-time feedback and adaptive interventions, this work represents a significant step toward creating dynamic and responsive learning environments that cater to individual needs.

5.3 Limitations and Future Work

The presented work has certain limitations, leaving scope for improvement and opportunities for further research in the following areas.

Sample Size and Generalizability: A common limitation across the studies is the relatively small and specific sample sizes, which may limit the generalizability of the findings to broader populations. Participants often represent a narrow demographic, such as university students, which may not account for variations in age, cultural backgrounds, educational levels, or cognitive abilities. Future studies should aim to include larger and more diverse participant pools to improve external validity and ensure the findings are applicable across different user groups.

Data Quality and Sensor Limitations: The quality and reliability of the data collected from sensors, such as gaze trackers and physiological devices, present challenges. Issues such as low sampling frequencies, calibration errors, noise, and inconsistencies in the data can impact the accuracy of predictions and model performance. For example, physiological data from devices like the Empatica E4 may not fully capture the complexity of cognitive or affective states. Future work should explore the use of more advanced sensors with higher sampling rates and improved calibration techniques to enhance data quality.

Feature Integration and Model Performance: The integration of multimodal features, such as gaze, physiological signals, and text, has not consistently yielded the expected improvements in model performance. This may be due to redundant or irrelevant information or insufficient preprocessing. Additionally, the deep learning models used in these studies often require extensive computational resources and large datasets, which can be a limitation when working with smaller or imbalanced datasets. Future research should focus on optimizing feature selection, exploring alternative model architectures, and addressing data imbalances to improve performance.

Generative AI for Adaptive Interventions: While the system effectively uses real-time gaze-based engagement detection and LLMs to generate summaries when engagement is low, there is room to expand the role of generative AI in adaptive learning. Future research could explore how GenAI can provide more diverse interventions, such as rephrased explanations, interactive quizzes, or visual aids, tailored to the learner's needs. Integrating multimodal data (e.g., gaze, facial expressions, and physiological signals) with GenAI could enable more context-aware and personalized interventions. Ensuring the accuracy, fairness, and explainability of AI-generated content remains a critical challenge that must be addressed to build trust and maximize learning outcomes.

Real-Time Feedback and Scalability: While the system provides real-time feedback, scaling it to larger and more diverse learning environments presents challenges. Future work should investigate how the system can handle increased user diversity, including variations in reading habits, engagement patterns, and content complexity.

Short-Term Focus and Long-Term Effects: Most studies focus on short-term tasks and immediate outcomes, limiting insights into long-term engagement, learning retention, and behavioral patterns. The lack of longitudinal studies makes it difficult to assess the impact of adaptive interventions on learning outcomes. Future research should explore the long-term effects of these systems, including their influence on knowledge retention, motivation, and overall learning effectiveness.

Ethical and Practical Considerations: The use of real-time data, such as gaze tracking and physiological signals, raises ethical concerns related to privacy, transparency, and user consent. The reliance on AI-generated content, such as summaries, introduces risks of oversimplification, inaccuracies, or biases in the information provided. Ensuring data security, explainability of AI-driven decisions, and the quality of AI-generated outputs are challenges that need to be addressed in future work.

Bibliography

[1] A. Abdul, C. von der Weth, M. Kankanhalli, and B. Y. Lim, “Cogam: Measuring and moderating cognitive load in machine learning model explanations,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[2] J. W. Ahn, Y. Ku, and H. C. Kim, “A novel wearable eeg and ecg recording system for stress assessment,” *Sensors*, vol. 19, no. 9, p. 1991, 2019.

[3] A. Almusaed, A. Almssad, I. Yitmen, and R. Z. Homod, “Enhancing student engagement: Harnessing “aied”’s power in hybrid education—a review analysis,” *Education Sciences*, vol. 13, no. 7, p. 632, 2023.

[4] 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 *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*, 2017, pp. 26–32.

[5] S. Amin, M. I. Uddin, A. A. Alarood, W. K. Mashwani, A. Alzahrani, and A. O. Alzahrani, “Smart e-learning framework for personalized adaptive learning and sequential path recommendations using reinforcement learning,” *IEEE Access*, 2023.

[6] 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,” *Scientific Reports*, vol. 12, no. 1, p. 5857, 2022.

[7] J. J. Appleton, S. L. Christenson, D. Kim, and A. L. Reschly, “Measuring cognitive and psychological engagement: Validation of the student engagement instrument,” *Journal of school psychology*, vol. 44, no. 5, pp. 427–445, 2006.

[8] 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*, Ios Press, 2009, pp. 17–24.

[9] I. Arroyo, B. P. Woolf, W. Burelson, K. Muldner, D. Rai, and M. Tai, “A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect,” *International Journal of Artificial Intelligence in Education*, vol. 24, pp. 387–426, 2014.

[10] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias, “Estimation of behavioral user state based on eye gaze and head pose—application in an e-learning environment,” *Multimedia Tools and Applications*, vol. 41, pp. 469–493, 2009.

[11] I. Azaiz, N. Kiesler, and S. Strickroth, "Feedback-generation for programming exercises with gpt-4," *arXiv preprint arXiv:2403.04449*, 2024.

[12] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*, IEEE, 2017, pp. 1–5.

[13] R. Balyan, K. S. McCarthy, and D. S. McNamara, "Applying natural language processing and hierarchical machine learning approaches to text difficulty classification," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 3, pp. 337–370, 2020.

[14] R. Biedert, G. Buscher, S. Schwarz, M. Möller, A. Dengel, and T. Lottermann, "The text 2.0 framework: Writing web-based gaze-controlled realtime applications quickly and easily," Feb. 2010. DOI: `10.1145/2002333.2002351`.

[15] A. T. Bimba, N. Idris, A. Al-Hunaiyyan, R. B. Mahmud, and N. L. B. M. Shuib, "Adaptive feedback in computer-based learning environments: A review," *Adaptive Behavior*, vol. 25, no. 5, pp. 217–234, 2017.

[16] R. Bixler and S. D'Mello, "Automatic gaze-based user-independent detection of mind wandering during computerized reading," *User Modeling and User-Adapted Interaction*, vol. 26, pp. 33–68, 2016.

[17] R. N. Blasiman, J. Dunlosky, and K. A. Rawson, "The what, how much, and when of study strategies: Comparing intended versus actual study behaviour," *Memory*, vol. 25, no. 6, pp. 784–792, 2017.

[18] P. Blondé, M. Sperduti, D. Makowski, and P. Piolino, "Bored, distracted, and forgetful: The impact of mind wandering and boredom on memory encoding," *Quarterly Journal of Experimental Psychology (2006)*, p. 174 702 182 110 263, Jun. 2021. DOI: `10.1177/17470218211026301`.

[19] B. S. Bloom, "Learning for mastery. instruction and curriculum. regional education laboratory for the carolinas and virginia, topical papers and reprints, number 1.," *Evaluation comment*, vol. 1, no. 2, n2, 1968.

[20] E. Bonner, K. Garvey, M. Miner, S. Godin, and H. Reinders, "Measuring real-time learner engagement in the japanese efl classroom," *Innovation in Language Learning and Teaching*, vol. 17, no. 2, pp. 254–264, 2023.

[21] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[22] P. J. Brantley, C. D. Waggoner, G. N. Jones, and N. B. Rappaport, "A daily stress inventory: Development, reliability, and validity," *Journal of behavioral medicine*, vol. 10, pp. 61–73, 1987.

[23] D. Brassell and T. Rasinski, *Comprehension that works: Taking students beyond ordinary understanding to deep comprehension*. Teacher Created Materials, 2008.

[24] 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, 2020, ISSN: 1424-8220. DOI: `10.3390/s20092546`. [Online]. Available: `http://dx.doi.org/10.3390/s20092546`.

[25] G. Buscher, A. Dengel, R. Biedert, and L. V. Elst, "Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 1, no. 2, pp. 1–30, 2012.

[26] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[27] M. Carroll et al., "Automatic detection of learner engagement using machine learning and wearable sensors," *Journal of Behavioral and Brain Science*, vol. 10, no. 3, pp. 165–178, 2020.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[29] J. Chen and P. B. D, "Construction of a learning engagement evaluation model based on multi-modal data fusion," in *2022 International Conference on Knowledge Engineering and Communication Systems (ICKES)*, 2022, pp. 1–7. DOI: `10.1109/ICKECS56523.2022.10060326`.

[30] Y. Chen, J. Zhou, Q. Gao, J. Gao, and W. Zhang, "Mdnn: Predicting student engagement via gaze direction and facial expression in collaborative learning.," *CMES-Computer Modeling in Engineering & Sciences*, vol. 136, no. 1, 2023.

[31] V. Chheang et al., "Towards anatomy education with generative ai-based virtual assistants in immersive virtual reality environments," in *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, IEEE, 2024, pp. 21–30.

[32] Y. Cho, S. J. Julier, and N. Bianchi-Berthouze, "Instant stress: Detection of perceived mental stress through smartphone photoplethysmography and thermal imaging," *JMIR mental health*, vol. 6, no. 4, e10140, 2019.

[33] S Cohen, T Kamarck, and R Mermelstein, *A global measure of perceived stress, journal of health and social behavior, vol. 24*, 1983.

[34] C. Conati, N. Jaques, and M. Muir, "Understanding attention to adaptive hints in educational games: An eye-tracking study," *International Journal of Artificial Intelligence in Education*, vol. 23, no. 1-4, pp. 136–161, 2013.

[35] L. Dabbish, G. Mark, and V. M. González, "Why do i keep interrupting myself? environment, habit and self-interruption," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 3127–3130.

[36] R. Deng and Y. Gao, "A review of eye tracking research on video-based learning," *Education and Information Technologies*, vol. 28, no. 6, pp. 7671–7702, 2023.

[37] L. R. Derogatis, *Brief symptom inventory: BSI*. Pearson, 1993.

[38] F. D'Errico, M. Paciello, and L. Cerniglia, "When emotions enhance students' engagement in e-learning processes," *Journal of e-Learning and Knowledge Society*, vol. 12, no. 4, 2016.

[39] F. S. Dhabhar, "The short-term stress response–mother nature's mechanism for enhancing protection and performance under conditions of threat, challenge, and opportunity," *Frontiers in neuroendocrinology*, vol. 49, pp. 175–192, 2018.

[40] Z Dharmawan and L. Rothkrantz, "Analysis of computer game player stress level using eeg data," in *11th international conference on computer games: AI, animation, mobile, educational and serious games, La Rochelle, France*, The University of Wolverhampton, 2007, pp. 111–124.

[41] 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, 2012.

[42] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM computing surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[43] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM computing surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[44] S. Dutta, S. Ranjan, S. Mishra, V. Sharma, P. Hewage, and C. Iwendi, "Enhancing educational adaptability: A review and analysis of ai-driven adaptive learning platforms," in *2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM)*, IEEE, 2024, pp. 1–5.

[45] 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, 2020.

[46] 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, 2020.

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

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

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

[50] H. A. El-Sabagh, "Adaptive e-learning environment based on learning styles and its impact on development students' engagement," *International Journal of Educational Technology in Higher Education*, vol. 18, no. 1, p. 53, 2021.

[51] S. G. Essa, T. Celik, and N. E. Human-Hendricks, "Personalized adaptive learning technologies based on machine learning techniques to identify learning styles: A systematic literature review," *IEEE Access*, vol. 11, pp. 48 392–48 409, 2023.

[52] A. Ezzaim, A. Dahbi, A. Haidine, and A. Aqqal, "Enhancing academic outcomes through an adaptive learning framework utilizing a novel machine learning-based performance prediction method," *Data and Metadata*, vol. 2, pp. 164–164, 2023.

[53] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad, "A comparison of features for automatic readability assessment," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, ser. COLING '10, Beijing, China: Association for Computational Linguistics, 2010, pp. 276–284. [Online]. Available: `http://dl.acm.org/citation.cfm?id=1944566.1944598`.

[54] R. Ferguson, "Learning analytics: Drivers, developments and challenges," *International Journal of Technology Enhanced Learning*, vol. 4, no. 5-6, pp. 304–317, 2012.

[55] R. Flesch, "A new readability yardstick.," *Journal of Applied Psychology*, vol. 32, no. 3, p221 –233, 1948, ISSN: 0021-9010. [Online]. Available: `http://libezproxy.open.ac.uk/login?url=http://search.ebscohost.com.libezproxy.open.ac.uk/login.aspx?direct=true&db=pdh&AN=apl-32-3-221&site=ehost-live&scope=site`.

[56] S. Forster, "Distraction and mind-wandering under load," *Frontiers in Psychology*, vol. 4, p. 283, 2013, ISSN: 1664-1078. DOI: `10.3389/fpsyg.2013.00283`. [Online]. Available: `https://www.frontiersin.org/article/10.3389/fpsyg.2013.00283`.

[57] H. Gabbay and A. Cohen, "Combining llm-generated and test-based feedback in a mooc for programming," in *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, 2024, pp. 177–187.

[58] P. Garg, J. Santhosh, A. Dengel, and S. Ishimaru, "Stress detection by machine learning and wearable sensors," in *26th International Conference on Intelligent User Interfaces-Companion*, 2021, pp. 43–45.

[59] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE transactions on affective computing*, vol. 13, no. 1, pp. 440–460, 2019.

[60] D. Girardi, N. Novielli, D. Fucci, and F. Lanubile, "Recognizing developers' emotions while programming," in *Proceedings of the ACM/IEEE 42nd international conference on software engineering*, 2020, pp. 666–677.

[61] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *Journal of biomedical informatics*, vol. 73, pp. 159–170, 2017.

[62] I. Gligorea, M. Cioca, R. Oancea, A.-T. Gorski, H. Gorski, and P. Tudorache, "Adaptive learning using artificial intelligence in e-learning: A literature review," *Education Sciences*, vol. 13, no. 12, p. 1216, 2023.

[63] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11 365–11 394, 2023.

[64] I. Gutiérrez, V. Álvarez, M. P. Paule, J. R. Pérez-Pérez, and S. De Freitas, "Adaptation in e-learning content specifications with dynamic sharable objects," *Systems*, vol. 4, no. 2, p. 24, 2016.

[65] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[66] 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, 2020.

[67] J. Harackiewicz and C. Hulleman, "The importance of interest: The role of achievement goals and task values in promoting the development of interest," *Social and Personality Psychology Compass*, vol. 4, pp. 42 –52, Dec. 2009. DOI: `10.1111/j.1751-9004.2009.00207.x`.

[68] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, Sage publications Sage CA: Los Angeles, CA, vol. 50, 2006, pp. 904–908.

[69] S. Harvey and A. Goudvis, *Strategies that work: Teaching comprehension for understanding and engagement*. Stenhouse publishers, 2007.

[70] 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 *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 73–80.

[71] 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 *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 73–80.

[72] C. E. Heil and D. F. Walnut, “Continuous and discrete wavelet transforms,” *SIAM review*, vol. 31, no. 4, pp. 628–666, 1989.

[73] M. J. Heilman, K. Collins, and J. Callan, “Combining lexical and grammatical features to improve readability measures for first and second language texts,” in *Proceedings of the Human Language Technology Conference*, 2007.

[74] W. S. Helton and K. Näswall, “Short stress state questionnaire,” *European Journal of Psychological Assessment*, 2015.

[75] C. R. Henrie, L. R. Halverson, and C. R. Graham, “Measuring student engagement in technology-mediated learning: A review,” *Computers & Education*, vol. 90, pp. 36–53, 2015.

[76] E. Heyd-Metzuyanim, E. S. Haataja, M. S. Hannula, and E. Garcia Moreno-Esteva, “What can eye-tracking, combined with discourse analysis, teach us about the ineffectiveness of a group of students solving a geometric problem?” *Instructional Science*, vol. 51, no. 3, pp. 363–396, 2023.

[77] G. Hinton, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[78] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. oup Oxford, 2011.

[79] A. G. Howard, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.

[80] K. Huang, T. Bryant, and B. Schneider, “Identifying collaborative learning states using unsupervised machine learning on eye-tracking, physiological and motion sensor data.,” *International Educational Data Mining Society*, 2019.

[81] T. Hussain, L. Yu, M. Asim, A. Ahmed, and M. A. Wani, “Enhancing e-learning adaptability with automated learning style identification and sentiment analysis: A hybrid deep learning approach for smart education,” *Information*, vol. 15, no. 5, p. 277, 2024.

[82] G.-J. Hwang, H. Xie, B. W. Wah, and D. Gašević, *Vision, challenges, roles and research issues of artificial intelligence in education*, 2020.

[83] S. Ishimaru et al., “Cognitive state measurement on learning materials by utilizing eye tracker and thermal camera,” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 08, 2017, pp. 32–36. DOI: 10.1109/ICDAR.2017.378.

[84] S. Ishimaru, S. S. Bukhari, C. Heisel, J. Kuhn, and A. Dengel, “Towards an intelligent textbook: Eye gaze based attention extraction on materials for learning and instruction in physics,” in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*, 2016, pp. 1041–1045.

[85] S. Ishimaru et al., "Cognitive state measurement on learning materials by utilizing eye tracker and thermal camera," in *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, IEEE, vol. 8, 2017, pp. 32–36.

[86] S. Ishimaru et al., "Augmented learning on anticipating textbooks with eye tracking," *Positive Learning in the Age of Information: A Blessing or a Curse?*, pp. 387–398, 2018.

[87] B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.

[88] S. Jacob, S. Ishimaru, S. S. Bukhari, and A. Dengel, "Gaze-based interest detection on newspaper articles," in *Proceedings of the 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction*, 2018, pp. 1–7.

[89] S. Jacob, S. Ishimaru, and A. Dengel, "Interest detection while reading newspaper articles by utilizing a physiological sensing wristband," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018, pp. 78–81.

[90] S. Jacob, S. Ishimaru, and A. Dengel, "Interest detection while reading newspaper articles by utilizing a physiological sensing wristband," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018, pp. 78–81.

[91] N. Jaques, C. Conati, J. M. Harley, and R. Azevedo, "Predicting affect from gaze data during interaction with an intelligent tutoring system," in *Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings 12*, Springer, 2014, pp. 29–38.

[92] J. Jeon and S. Lee, "Large language models in education: A focus on the complementary relationship between human teachers and chatgpt," *Education and Information Technologies*, vol. 28, no. 12, pp. 15 873–15 892, 2023.

[93] X. Jin, S. Chai, J. Tang, X. Zhou, and K. Wang, "Eye-tracking in ar/vr: A technological review and future directions," *IEEE Open Journal on Immersive Displays*, 2024.

[94] N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th annual international symposium on computer architecture*, 2017, pp. 1–12.

[95] W. Kaiss, K. Mansouri, and F. Poirier, "Effectiveness of an adaptive learning chatbot on students' learning outcomes based on learning styles.," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 13, 2023.

[96] S. Karumbaiah, R. Lizarralde, D. Allessio, B. Woolf, I. Arroyo, and N. Wixon, "Addressing student behavior and affect with empathy and growth mindset.," *International Educational Data Mining Society*, 2017.

[97] F. Ke, R. Liu, Z. Sokolikj, I. Dahlstrom-Hakki, and M. Israel, "Using eye-tracking in education: Review of empirical research and technology," *Educational technology research and development*, pp. 1–36, 2024.

[98] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2017, pp. 985–990.

[99] M. A. Killingsworth and D. T. Gilbert, "A wandering mind is an unhappy mind," *Science*, vol. 330, no. 6006, pp. 932–932, 2010, ISSN: 0036-8075. DOI: `10.1126/science.1192439`. eprint: `https://science.sciencemag.org/content/330/6006/932.full.pdf`. [Online]. Available: `https://science.sciencemag.org/content/330/6006/932`.

[100] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[101] S. Koelstra et al., "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[102] A. C. Koenka and E. M. Anderman, "Personalized feedback as a strategy for improving motivation and performance among middle school students," *Middle School Journal*, vol. 50, no. 5, pp. 15–22, 2019.

[103] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," in *Proceedings of the 16th international conference on multimodal interaction*, 2014, pp. 291–298.

[104] A. Krapp, "Interest, learning and academic achievement. paper prepared for the symposium on" task motivation by interest," in *Third European Conference of Learning and Instruction (EARLI), Madrid, Spain, September*, 1989.

[105] A. Krapp, "The construct of interest: Characteristics of indvidual interests and interest–related actions from the perspective of a person–object–theory," *Studies in Educational Psychology*, vol. 1, Jan. 1993.

[106] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th international conference on intelligent user interfaces*, 2015, pp. 126–137.

[107] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563, 2017.

[108] F. Z. Lhafra and O. Abdoun, "Towards an adaptive learning process using artificial intelligence technologies," in *International Conference on Digital Technologies and Applications*, Springer, 2023, pp. 23–32.

[109] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition," *Information Processing & Management*, vol. 57, no. 3, p. 102 185, 2020.

[110] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, "An explainable deep fusion network for affect recognition using physiological signals," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 2069–2072.

[111] H.-L. Liu, T.-H. Wang, H.-C. K. Lin, C.-F. Lai, and Y.-M. Huang, "The influence of affective feedback adaptive learning system on learning engagement and self-directed learning," *Frontiers in Psychology*, vol. 13, p. 858 411, 2022.

[112] X. Long, S. Mayer, and F. Chiossi, "Multimodal detection of external and internal attention in virtual reality using eeg and eye tracking features," in *Proceedings of Mensch Und Computer 2024*, 2024, pp. 29–43.

[113] 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, 2017.

[114] B. Mahesh, T. Hassan, E. Prassler, and J.-U. Garbas, "Requirements for a reference dataset for multimodal human stress detection," in *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, IEEE, 2019, pp. 492–498.

[115] R. L. Mandryk and M. S. Atkins, "A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies," *International Journal of Human-Computer Studies*, vol. 65, no. 4, pp. 329–347, 2007.

[116] F. Martin, Y. Chen, R. L. Moore, and C. D. Westine, "Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018," *Educational Technology Research and Development*, vol. 68, no. 4, pp. 1903–1929, 2020.

[117] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface 05 audio-visual emotion database," Feb. 2006, pp. 8 –8, ISBN: 0-7695-2571-7. DOI: `10.1109/ICDEW.2006.145`.

[118] B. R. McCoy, "Digital distractions in the classroom phase ii: Student classroom use of digital devices for non-class related purposes.(2016)," *Faculty Publications, College of Journalism & Mass Communications.*, vol. 90, pp. 1–44, 2016.

[119] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2013, pp. 881–888.

[120] 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 *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, 2016, pp. 3723–3726.

[121] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, PMLR, 2017, pp. 1273–1282.

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

[123] 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 Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, 2018.

[124] M. Moisala et al., "Media multitasking is associated with distractibility and increased prefrontal activity in adolescents and young adults," *NeuroImage*, vol. 134, pp. 113–121, 2016.

[125] G.-A. Morosanu, L.-A. Rata, M. Geru, et al., "Outcomes of large language models and artificial intelligence in education," *Didactica Danubiensis*, vol. 3, no. 1, pp. 30–52, 2023.

[126] R. Murphy, "Artificial intelligence applications to support k-12 teachers and teaching," *The RAND corporation*, 2019.

[127] R. Mutegeki and D. S. Han, "A cnn-lstm approach to human activity recognition," in *2020 international conference on artificial intelligence in information and communication (ICAIIC)*, IEEE, 2020, pp. 362–366.

[128] S Narciss, "How to design informative tutoring feedback for multimedia learning," *Instructional Design for Meltimedia Learning/Waxmann*, 2004.

[129] S. Narciss et al., "Exploring feedback and student characteristics relevant for personalizing feedback strategies," *Computers & Education*, vol. 71, pp. 56–76, 2014.

[130] V. Navalpakkam, J. M. Rao, and M. Slaney, "Using gaze patterns to study and predict reading struggles due to distraction," in *CHI EA '11*, 2011.

[131] K. Nkurikiyeyezu, K. Shoji, A. Yokokubo, and G. Lopez, "Thermal comfort and stress recognition in office environment.," in *HEALTHINF*, 2019, pp. 256–263.

[132] N. B. A. Normadhi, L. Shuib, H. N. M. Nasir, A. Bimba, N. Idris, and V. Balakrishnan, "Identification of personal traits in adaptive learning environment: Systematic literature review," *Computers & Education*, vol. 130, pp. 168–190, 2019.

[133] H. S. Nwana, "Intelligent tutoring systems: An overview," *Artificial Intelligence Review*, vol. 4, no. 4, pp. 251–277, 1990.

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

[135] A. Omar, K. Hassane, and K. Mohamed, "Towards a new generation of intelligent, adaptive e-learning platforms," *International Journal of Computing and Digital Systems*, vol. 17, no. 1, pp. 1–10, 2024.

[136] E. Ophir, C. Nass, and A. D. Wagner, "From the cover: Cognitive control in media multitaskers," *Proceedings of the national academy of sciences of the United States of America*, vol. 106, no. 37, p. 15 583, 2009.

[137] F. Ouyang and P. Jiao, "Artificial intelligence in education: The three paradigms," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100 020, 2021.

[138] A. P. Pai, J. Santhosh, and S. Ishimaru, "Real-time feedback on reader's engagement and emotion estimated by eye-tracking and physiological sensing," in *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, 2022, pp. 97–98.

[139] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[140] A. Pardo, J. Jovanovic, S. Dawson, D. Gašević, and N. Mirriahi, "Using learning analytics to scale the provision of personalised feedback," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 128–138, 2019.

[141] J. Parsons and L. Taylor, "Improving student engagement," *Current issues in education*, vol. 14, no. 1, 2011.

[142] M. Perkins, "Academic integrity considerations of ai large language models in the post-pandemic era: Chatgpt and beyond," *Journal of University Teaching and Learning Practice*, vol. 20, no. 2, 2023.

[143] R. W. Picard, "Affective computing: Challenges," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003.

[144] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[145] T. G. Pickering et al., "Environmental influences on blood pressure and the role of job strain.," *Journal of hypertension. Supplement: official journal of the International Society of Hypertension*, vol. 14, no. 5, S179–85, 1996.

[146] M.-Z. Poh, N. C. Swenson, and R. W. Picard, "A wearable sensor for unobtrusive, long-term assessment of electrodermal activity," *IEEE transactions on Biomedical engineering*, vol. 57, no. 5, pp. 1243–1252, 2010.

[147] J.-L. Qiu, X.-Y. Li, and K. Hu, "Correlated attention networks for multimodal emotion recognition," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 2656–2660. DOI: 10.1109/BIBM.2018.8621129.

[148] R. A. Ramachandra, J. Santhosh, A. Dengel, and S. Ishimaru, "Enhancing stress detection for students: Exploring the impact of fine-tuning and user-specific data calibration in deep learning," *International Journal of Activity and Behavior Computing*, vol. 2024, no. 1, pp. 1–21, 2024.

[149] K. Rayner, K. Chace, T. Slattery, and J. Ashby, "Eye movements as reflections of comprehension processes in reading," *Scientific Studies of Reading - SCI STUD READ*, vol. 10, Jul. 2006. DOI: 10.1207/s1532799xssr1003_3.

[150] K. Renninger and R. Wozniak, "Effect of interest on attentional shift, recognition, and recall in young children," *Developmental Psychology*, vol. 21, pp. 624–632, Jul. 1985. DOI: 10.1037/0012-1649.21.4.624.

[151] C. R. Rogers, "Freedom to leamjbr the 80's," 1983.

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

[153] J. I. Rotgans and H. G. Schmidt, "Situational interest and academic achievement in the active-learning classroom," *Learning and Instruction*, vol. 21, no. 1, pp. 58–67, 2011, ISSN: 0959-4752. DOI: https://doi.org/10.1016/j.learninstruc.2009.11.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095947520900111X.

[154] B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Jan. 2019.

[155] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, 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, 2018.

[156] J. Santhosh, A. Dengel, and S. Ishimaru, "Gaze-driven adaptive learning system with chatgpt-generated summaries," *IEEE Access*, vol. 12, pp. 173 714–173 733, 2024.

[157] J. Santhosh, D. Dzsotjan, and S. Ishimaru, “Multimodal assessment of interest levels in reading: Integrating eye-tracking and physiological sensing,” *IEEE Access*, 2023.

[158] J. Santhosh, S. Ishimaru, and A. Dengel, “Generating heatmap for unknown documents towards readability measurement,” in *Companion Proceedings of the 25th International Conference on Intelligent User Interfaces*, 2020, pp. 47–48.

[159] J. Santhosh, S. Jagtap, A. Dengel, and S. Ishimaru, “Digital distractions in reading: Investigating impact of cognitive control training on reading behavior and outcomes,” in *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2024, pp. 91–95.

[160] J. Santhosh, A. P. Pai, and S. Ishimaru, “Toward an interactive reading experience: Deep learning insights and visual narratives of engagement and emotion,” *IEEE Access*, 2024.

[161] T. Sargazi Moghadam, A. Darejeh, M. Delaramifar, and S. Mashayekh, “Toward an artificial intelligence-based decision framework for developing adaptive e-learning systems to impact learners’ emotions,” *Interactive Learning Environments*, pp. 1–21, 2023.

[162] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, “Introducing wesad, a multimodal dataset for wearable stress and affect detection,” in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, pp. 400–408.

[163] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, “Introducing wesad, a multimodal dataset for wearable stress and affect detection,” in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, pp. 400–408.

[164] P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven, “Wearable-based affect recognition—a review,” *Sensors*, vol. 19, no. 19, p. 4079, 2019.

[165] S. E. Schwarm and M. Ostendorf, “Reading level assessment using support vector machines and statistical language models,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL ’05, Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 523–530. DOI: `10.3115/1219840.1219905`. [Online]. Available: `https://doi.org/10.3115/1219840.1219905`.

[166] 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 *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[167] R. Shadiev and D. Li, “A review study on eye-tracking technology usage in immersive virtual reality learning environments,” *Computers & education*, vol. 196, p. 104 681, 2023.

[168] S. Shaikh et al., “Emovis–an interactive visualization tool to track emotional trends during crisis events,” in *Advances in Artificial Intelligence, Software and Systems Engineering: Proceedings of the AHFE 2019 International Conference on Human Factors in Artificial Intelligence and Social Computing, the AHFE International Conference on Human Factors, Software, Service and Systems Engineering, and the AHFE International Conference of Human Factors in Energy, July 24-28, 2019, Washington DC, USA 10*, Springer, 2020, pp. 14–24.

[169] P. Sharma et al., “Student engagement detection using emotion analysis, eye tracking and head movement with machine learning,” in *International Conference on Technology and Innovation in Learning, Teaching and Education*, Springer, 2022, pp. 52–68.

[170] D. J. Shernoff, “Optimal learning environments to promote student engagement,” 2013.

[171] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge computing: Vision and challenges,” *IEEE internet of things journal*, vol. 3, no. 5, pp. 637–646, 2016.

[172] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. Rahaman, and T. Zia, “Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals,” *Journal of Network and Computer Applications*, vol. 149, p. 102 447, 2020.

[173] V. Shute and B. Towle, “Adaptive e-learning,” in *Aptitude*, Routledge, 2018, pp. 105–114.

[174] R. da Silva Soares Jr, C. Barreto, and J. R. Sato, “Perspectives in eye-tracking technology for applications in education,” *South African Journal of Childhood Education*, vol. 13, no. 1, p. 1204, 2023.

[175] P. Skapinakis, “Spielberger state-trait anxiety inventory,” in *Encyclopedia of Quality of Life and Well-Being Research*, A. C. Michalos, Ed. Dordrecht: Springer Netherlands, 2014, pp. 6261–6264, ISBN: 978-94-007-0753-5. DOI: 10.1007/978-94-007-0753-5_2825. [Online]. Available: https://doi.org/10.1007/978-94-007-0753-5_2825.

[176] V. Skaramagkas et al., “Review of eye tracking metrics involved in emotional and cognitive processes,” *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 260–277, 2021.

[177] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A multimodal database for affect recognition and implicit tagging,” *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 42–55, 2011.

[178] J.-M. Su, S.-S. Tseng, H.-Y. Lin, and C.-H. Chen, “A personalized learning content adaptation mechanism to meet diverse user needs in mobile learning environments,” *User modeling and user-adapted interaction*, vol. 21, pp. 5–49, 2011.

[179] Y.-T. Sung, J.-L. Chen, J.-H. Cha, H.-C. Tseng, T.-H. Chang, and K.-E. Chang, "Constructing and validating readability models: The method of integrating multilevel linguistic features with machine learning," *Behavior research methods*, vol. 47, pp. 340–354, 2015.

[180] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[181] M. Taub, R. Azevedo, F. Bouchet, and B. Khosravifar, "Can the use of cognitive and metacognitive self-regulated learning strategies be predicted by learners' levels of prior knowledge in hypermedia-learning environments?" *Computers in Human Behavior*, vol. 39, pp. 356–367, 2014.

[182] L. Tetzlaff, F. Schmiedek, and G. Brod, "Developing personalized education: A dynamic framework," *Educational Psychology Review*, vol. 33, pp. 863–882, 2021.

[183] V. Tran, "Positive affect negative affect scale (panas)," in *Encyclopedia of Behavioral Medicine*, M. D. Gellman and J. R. Turner, Eds. New York, NY: Springer New York, 2013, pp. 1508–1509, ISBN: 978-1-4419-1005-9. DOI: 10.1007/978-1-4419-1005-9_978. [Online]. Available: https://doi.org/10.1007/978-1-4419-1005-9_978.

[184] D. Turnbull, R. Chugh, and J. Luck, "Learning management systems, an overview," *Encyclopedia of education and information technologies*, pp. 1052–1058, 2020.

[185] C. M. Tyng, H. U. Amin, M. N. Saad, and A. S. Malik, "The influences of emotion on learning and memory," *Frontiers in psychology*, vol. 8, p. 235 933, 2017.

[186] P. Tzirakis, G. Trigeorgis, M. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 1301–1309, 2017. DOI: 10.1109/JSTSP.2017.2764438.

[187] P. Vadiraja, J. Santhosh, H. Moulay, A. Dengel, and S. Ishimaru, "Effects of counting seconds in the mind while reading," in *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, 2021, pp. 486–490.

[188] G. Verma and U. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage*, vol. 102, pp. 162–172, 2014. DOI: https://doi.org/10.1016/j.neuroimage.2013.11.007.

[189] J. C. Wallace and S. J. Vodanovich, "Can accidents and industrial mishaps be predicted? further investigation into the relationship between cognitive failure and reports of accidents," *Journal of Business and Psychology*, vol. 17, no. 4, pp. 503–514, 2003.

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

[191] 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 Sens. J.*, vol. 21, pp. 22 402–22 413, 2021.

[192] Y. Wang et al., "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83, pp. 19–52, 2022.

[193] 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. 52 886–52 898, 2023. DOI: 10.1109/ACCESS.2023.3279428.

[194] H. Wei, A.-N. Moldovan, and C. H. Muntean, "Sensing learner interest through eye tracking," 2009.

[195] D. Wu, J. Zhang, and Q. Zhao, "Multimodal fused emotion recognition about expression-eeg interaction and collaboration using deep learning," *IEEE Access*, vol. PP, pp. 1–1, Jul. 2020. DOI: 10.1109/ACCESS.2020.3010311.

[196] G. Wu, G. Liu, and M. Hao, "The analysis of emotion recognition from gsr based on pso," in *2010 International symposium on intelligence information processing and trusted computing*, IEEE, 2010, pp. 360–363.

[197] E. Yang et al., "Adatask: A task-aware adaptive learning rate approach to multi-task learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, 2023, pp. 10 745–10 753.

[198] S.-N. Yang and G. Mcconkie, "Eye movements during reading: A theory of saccade initiation times," *Vision research*, vol. 41, pp. 3567–85, Feb. 2001. DOI: 10.1016/S0042-6989(01)00025-6.

[199] C.-B. Yao, "Constructing a user-friendly and smart ubiquitous personalized learning environment by using a context-aware mechanism," *IEEE transactions on learning technologies*, vol. 10, no. 1, pp. 104–114, 2015.

[200] L. Yardley et al., "Understanding and promoting effective engagement with digital behavior change interventions," *American journal of preventive medicine*, vol. 51, no. 5, pp. 833–842, 2016.

[201] 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?" *International journal of environmental research and public health*, vol. 17, no. 7, p. 2320, 2020.

[202] D. Zhang and J. F. Nunamaker, "Powering e-learning in the new millennium: An overview of e-learning and enabling technology," *Information systems frontiers*, vol. 5, pp. 207–218, 2003.

[203] F. Zhang, X. Wang, and X. Zhang, "Applications of deep learning method of artificial intelligence in education," *Education and Information Technologies*, pp. 1–25, 2024.

[204] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.

[205] X. Zhang et al., "Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine," *IEEE transactions on cybernetics*, vol. 51, no. 9, pp. 4386–4399, 2020.

[206] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification," *Journal of systems engineering and electronics*, vol. 28, no. 1, pp. 162–169, 2017.

[207] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification," *Journal of Systems Engineering and Electronics*, vol. 28, no. 1, pp. 162–169, 2017.

[208] W. X. Zhao et al., "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

Own Publications

Journal Papers

1. Jayasankar Santhosh, Andreas Dengel and Shoya Ishimaru. “Gaze-Driven Adaptive Learning System with ChatGPT-Generated Summaries”. IEEE Access, pp. 173714–173733, 2024.

2. Jayasankar Santhosh, Akshay Palimar Pai and Shoya Ishimaru. “Toward an Interactive Reading Experience: Deep Learning Insights and Visual Narratives of Engagement and Emotion”. IEEE Access, pp. 6001–6016, 2024.

3. Jayasankar Santhosh, David Dzsotjan and Shoya Ishimaru. “Multimodal Assessment of Interest Levels in Reading: Integrating Eye-Tracking and Physiological Sensing”. IEEE Access, pp. 93994–94008, 2023.

4. Rashmi Alur Ramachandra, Jayasankar Santhosh, Andreas Dengel and Shoya Ishimaru. “Enhancing Stress Detection for Students: Exploring the Impact of Fine-Tuning and User-Specific Data Calibration in Deep Learning”. International Journal of Activity and Behavior Computing 2024 (1), pp. 1–15, 2024.

International Conference Papers

1. Jayasankar Santhosh, Shrishti Jagtap, Andreas Dengel, and Shoya Ishimaru. 2024. Digital Distractions in Reading: Investigating Impact of Cognitive Control Training on Reading Behavior and Outcomes. In Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp ’24). Association for Computing Machinery, New York, NY, USA, 91–95, 2024.

2. Akshay Palimar Pai, Jayasankar Santhosh, Shoya Ishimaru. "Real-Time Feedback on Reader’s Engagement and Emotion Estimated by Eye-Tracking and Physiological Sensing”. In Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp ’22 Adjunct), pp. 97–98, 2022.

3. Pramod Vadiraja, Jayasankar Santhosh, Hanane Moulay, Andreas Dengel and Shoya Ishimaru. “Effects of Counting Seconds in the Mind while Reading”. In Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp ’21 Adjunct), pp. 486–490, 2021.

4. Prerna Garg, Jayasankar Santhosh, Andreas Dengel and Shoya Ishimaru. “Stress Detection by Machine Learning and Wearable Sensors”. In Proceedings of the 26th International Conference on Intelligent User Interfaces Companion (IUI ’21), pp. 237–240, 2021

5. Jayasankar Santhosh, Shoya Ishimaru and Andreas Dengel. “Generating Heatmap for Unknown Documents Towards Readability Measurement”. In Proceedings of the 25th International Conference on Intelligent User Interfaces Companion (IUI ’20), pp. 47–48, 2020.

6. Jayasankar Santhosh, Shoya Ishimaru and Andreas Dengel. “Estimating Fixation Durations for Each Word in Documents towards Readability Measurement”. In Proceedings of the 1st international interdisciplinary Symposium on Reading Experience and Analysis of Documents (READ ’18), 2018.

Papers Not Related to the Thesis or with My Small Contribution

1. Ankur Bhatt, Ko Watanabe, Jayasankar Santhosh, Andreas Dengel and Shoya Ishimaru. “Estimating Self-Confidence in Video-Based Learning Using Eye-Tracking and Deep Neural Networks”. IEEE Access, pp. 192219–192229, 2024.

2. Ankur Bhatt, Ko Watanabe, Jayasankar Santhosh, Andreas Dengel, and Shoya Ishimaru. "ABCDE: Appearance-Based Confidence Detection by Evaluating Gaze Behavior Using Deep Learning." International Journal of Activity and Behavior Computing 2025, no. 1 (2025): 1-24.

3. Ko Watanabe, Gitesh Gund, Jayasankar Santhosh, Haruka Sakagami, Yuki Matsuda, Andreas Dengel, and Shoya Ishimaru. "EyeUnderstand: Dashboard for Gaze and Deep-Learning Driven Comprehension Estimation in Online Lectures." IEEE Access (2025).

Curriculum Vitae

Summary

Jayasankar Santhosh is a Postdoctoral Researcher at Zeppelin University, Friedrichshafen, working with the Chair of Artificial Intelligence. Prior to this, he was a Ph.D. Researcher at the German Research Center for Artificial Intelligence (DFKI), where he worked from 2019 to 2025. At DFKI, he was part of the Immersive Quantified Learning Laboratory (IQL Lab) and the Smart Data and Knowledge Services (SDS) Department in Kaiserslautern. In addition, he served as a Teaching Assistant at RPTU Kaiserslautern-Landau from 2022 to 2023.

He earned a bachelor's degree in Computer Science from Mahatma Gandhi University, India, in 2014, and a master's degree in Computer Science from RPTU Kaiserslautern-Landau, Germany, in 2018. He defended his Ph.D. in Engineering at RPTU Kaiserslautern-Landau in 2025. From 2017 to 2019, he worked as a Research Assistant at DFKI, where he developed scalable AI systems, optimized machine learning pipelines for document analysis, and applied advanced techniques to process unstructured text and sensor data.

His research interests include human-computer interaction, deep learning-based affective state recognition, AI for group productivity, sensor data analysis, and Generative AI (GenAI) for improving individual and group performance. He has published his work in venues such as IEEE Access, Ubicomp, IUI, and IJABC. He is also a Professional Member of the Association for Computing Machinery (ACM).

Background

Human-Computer Interaction, Generative AI for group productivity, AI and Learning Analytics, Eye-Tracking, Wearable Computing, Deep Learning and Time Series Analysis

Work Experience

Postdoctoral Researcher, Zeppelin University, Friedrichshafen — *2025 – present*

PhD Researcher, German Research Center for Artificial Intelligence (DFKI) — *2019 – 2025*

University Teaching Assistant, RPTU Kaiserslautern-Landau — *2022 – 2023*

Research Assistant, German Research Center for Artificial Intelligence (DFKI) — *2017 – 2019*

Education

Master of Science (Informatik), RPTU Kaiserslautern-Landau — *2015 – 2018*

Bachelor of Technology, Mahatma Gandhi University, Kerala — *2010 – 2014*

Professional Activities and Honors

Reviewed several scientific conference papers and journal papers(CHI, AH, Ubicomp, IUI), contributing to the advancement of research in AI.

Received Central Board of Secondary Education (CBSE) Merit Certificate for being in the top 0.1% in the AISSCE examination.

Selected Publications

Jayasankar Santhosh, Andreas Dengel and Shoya Ishimaru. "Gaze-Driven Adaptive Learning System With ChatGPT-Generated Summaries". IEEE Access, pp. 173714–173733, 2024. *Journal Paper*

Prerna Garg, Jayasankar Santhosh, Andreas Dengel and Shoya Ishimaru. "Stress Detection by Machine Learning and Wearable Sensors". In Proceedings of the 26th International Conference on Intelligent User Interfaces Companion (IUI '21), pp. 237–240, 2021 *Citation count of 114*