A Review of Multimodal Interaction in Remote Education: Technologies, Applications, and Challenges

Xie, Yangmei
Yang, Liuyi
Zhang, Miao
Chen, Sinan
Li, Jialong

(Citation)
Applied Sciences,15(7):3937

(Issue Date)
2025-04

(Resource Type)
journal article

(Version)
Version of Record

applied sciences

MDPI

Review

A Review of Multimodal Interaction in Remote Education: Technologies, Applications, and Challenges

Yangmei Xie [1], Liuyi Yang [2,*], Miao Zhang [3,*], Sinan Chen [4,5] and Jialong Li [6]

[1] Graduate School of Human Development and Environment, Kobe University, 3-11 Tsurukabe, Nada-ku, Kobe 657-8501, Japan; 247d605d@gsuite.kobe-u.ac.jp
[2] Graduate School of System Informatics, Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan
[3] Graduate School of Maritime Sciences, Kobe University, 5-1-1 Fukae Minamicho, Higashinada-ku, Kobe 658-0022, Japan
[4] Center of Mathematical and Data Sciences, Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan; chensinan@gold.kobe-u.ac.jp
[5] Graduate School of Engineering Faculty of Engineering, Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan
[6] Department of Computer Science and Engineering, Waseda University, 1-104 Totsukamachi, Shinjuku-ku, Tokyo 169-8050, Japan; lijialong@fuji.waseda.jp
* Correspondence: 211x508x@gsuite.kobe-u.ac.jp (L.Y.); 218w402w@gsuite.kobe-u.ac.jp (M.Z.)

Abstract: Multimodal interaction technology has become a key aspect of remote education by enriching student engagement and learning results as it utilizes the speech, gesture, and visual feedback as various sensory channels. This publication reflects on the latest breakthroughs in multimodal interaction and its usage in remote learning environments, including a multi-layered discussion that addresses various levels of learning and understanding. It showcases the main technologies, such as speech recognition, computer vision, and haptic feedback, that enable the visitors and learning portals to exchange data fluidly. In addition, we investigate the function of multimodal learning analytics in order to measure the cognitive and emotional states of students, targeting personalized feedback and refining instructional strategies. Though multimodal communication may bring a historical improvement to the mode of online education, the platform still faces many issues, such as media synchronization, higher computational demand, physical adaptability, and privacy concerns. These problems demand further research in the fields of algorithm optimization, access to technology guidance, and the ethical use of big data. This paper presents a systematic review of the application of multimodal interaction in remote education. Through the analysis of 25 selected research papers, this review explores key technologies, applications, and challenges in the field. By synthesizing existing findings, this study highlights the role of multimodal learning analytics, speech recognition, gesture-based interaction, and haptic feedback in enhancing remote learning.

Keywords: multimodality; education; remote learning technology; human–computer interaction; learning analytics

Academic Editor: Wonjoon Kim

Received: 28 February 2025
Revised: 21 March 2025
Accepted: 23 March 2025
Published: 3 April 2025

Citation: Xie, Y.; Yang, L.; Zhang, M.; Chen, S.; Li, J. A Review of Multimodal Interaction in Remote Education: Technologies, Applications, and Challenges. *Appl. Sci.* **2025**, *15*, 3937. https://doi.org/10.3390/app15073937

1. Introduction

The word "multimodal" defines the blending of more than one human sense modality, which includes vision, touch, hearing, and taste and smell, as well as communication skills like perception, cognition, and interaction. The main purpose of multimodal interaction is to help users better understand the information presented by computing systems [1]. This is usually achieved by technologies that are based on user senses, such as haptic devices, virtual reality (VR) systems, and augmented reality (AR) systems.

In systems of multimodal interaction, users can participate in interactions that are either uni-directional or bi-directional, which makes it easier to monitor and study user activities [2]. This technology is considered important for interactions with computers at the virtual level. As an example, Saffaryazdi et al. [2] studied verbal and non-verbal user behaviors, while engaging with virtual agents and supplemented emotion recognition through electroencephalography (EEG) and electrodermal activity (EDA) to improve the experience for users. Furthermore, some researchers have reported that visual and dynamic feedback [3] can enhance user engagement in virtual places more easily [4]. Geiger et al. [4] showed how virtual environments can be enhanced by optimizing visual feedback for virtual reality environments to improve instruction for skilled digital human–computer models, while exploiting the models for improved instructional outcomes of grasping actions.

Dynamic feedback systems are an essential feature of multimodal interaction. For instance, Chollet et al. [3] designed a system that allows users to give virtual speeches to an audience that can provide feedback as the user speaks through voices, gestures, and even facial expressions. This improves the skill of public speaking among users. In the same vein, Kotranza et al. [5] investigated haptic interaction on a virtual agent, enabling two-way communications by proving that virtual agents not only act on the user's actions, but also communicate through haptic response, thereby increasing user engagement.

The main goal of multimodal interaction is to make human–computer communication more natural, allowing systems to better understand users' intentions and emotions [6]. This technology is used not only in AR and VR but also in remote education, where it helps improve learning experiences and results. In recent years, the fast development of AI, computer vision, and sensors has led to the wide use of multimodal interaction in education. By combining speech, gestures, and facial expressions, multimodal feedback can effectively increase learners' engagement and understanding.

Traditional learning analytics mainly use single data sources, such as clickstreams, log records, and interaction data from learning management systems (LMS). These sources show only part of the learning process and do not provide a full understanding [7].

Learning is also shaped by hidden factors, such as cognitive states and emotional changes, which strongly affect learning outcomes [8]. Standard data collection methods often miss these behavioral, attentional, and emotional shifts. Multimodal data analysis can help address this gap.

Studies show that traditional learning analytics rely on limited data, like click data and system logs, which do not fully capture learning activities [9]. Multimodal learning analytics, however, combines different data types, such as EEG, EDA, facial expressions, and speech signals. This approach provides a clearer picture of learners' cognitive and emotional states, making personalized learning more effective [10].

Multimodal interaction technology offers a way to overcome the limits of traditional learning analytics. It makes remote education more flexible and engaging while giving learners real-time, personalized feedback. As AI, computer vision, and sensor technologies improve, multimodal interaction will play a larger role in education, offering new ways to personalize learning and assess students.

Although existing studies have preliminarily explored the application of multimodal interaction technologies in remote education (e.g., speech recognition, affective computing), the following research gaps remain unresolved:

- Real-time multimodal data fusion and latency mitigation in authentic classroom scenarios require more robust solutions;
- There is a lack of systematic research on the cross-cultural and multilingual adaptability of multimodal systems;

- Most empirical validations are conducted in controlled laboratory environments, with limited large-scale data from real-world educational settings.

This paper adopts the literature review method to review the representative research papers on educational multimodal interaction published in recent years and identify the technological advances, applications and challenges in this field.

2. Development and Application of Multimodal Interaction Technology in Remote Education

Before discussing the technical development of multimodal interaction, it is helpful to illustrate how such systems are typically structured in remote education settings. Figure 1 presents a general flow of information in a multimodal learning system. It begins with learner input, followed by data acquisition using sensor devices, and then proceeds to data processing and analysis. The system responds through adaptive feedback delivered to learners and instructors.

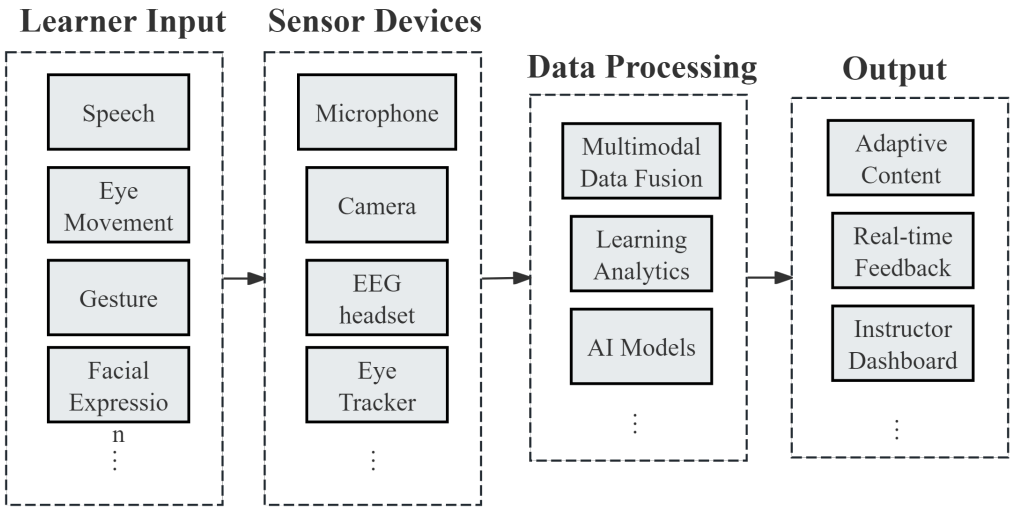

Figure 1. Flow of multimodal interaction in remote education.

2.1. Development of Multimodal Interaction Technology

This early research on multimodal interaction included the basics of speech and gesture recognition and proceeded to focus on developing real-world applications only for simple human–computer interactions [11]. However, today, the active use of artificial intelligence, deep learning, and sensor technologies enables bidirectional interaction, which, in turn, gives rise to real-time data fusion in multiple modalities. For example, the introduction of the technologies of emotion analytics, speech tone detection, and gesture interpretation brings quality and efficiency improvements in the human–computer interaction area (HCA) [12,13]. Additionally, the emergence of the Mixed Reality (MR) and AR has brought about the convenient and instant cooperation of visible, perceptual, and auditory information, and as a result, the qualitative and usability level of interactive systems has improved [14]. Recent technology emphasizes optimizing end-user experiences at its core, which particularly relates to facilitating accessibility to users who may have particular needs. For instance, moving multi-modal interaction (MMI) has generated beneficiaries in smart devices for the elderly, and it was observed that interaction efficiency was significantly enhanced [15]. In addition, analyses show that designers of interacting systems are centering on various user communities and progressively exploring how to customize holistic platforms to be irrefutably user-friendly [16].

2.2. Key Technologies in Remote Interaction

Speech- and Text-Based Interaction: Speech-based communicative modes, as well as natural language processing, are essential elements for a multi-channel interaction systems. These technologies have been integrated into different digital platforms, thus facilitating language-based interactions in a more natural and intuitive way [11,17].

Visual- and Gesture-Based Interaction: More recently, computer vision technology has been used with cameras to sense gestures as well as facial expressions, enabling the simultaneous recognition of the movements and emotions of the user in real time [13,14].

Haptic- and Motion-Based Interaction: Employing devices with tactile feedback, for instance, gloves that are sensitive to force and touch-sensitive interfaces, will improve users' sense of physical involvement in the interaction. Therefore, these technologies are extremely important in the field of smart manufacturing, remote collaboration, and online teaching [18].

Multimodal Data Fusion: Deep learning methods implement different fusion techniques, either early or late, or hybrid fusion models, in order to enrich speech, text, and visual data to generate multimodal data that in the end will lead to the development of advanced artificial intelligent systems with more elaborate context adaptations [19,20].

2.3. Applications of Multimodal Interaction Technology in Enhancing Educational Models

Learning is intrinsically multimodal since we use several sensory channels during communication and knowledge acquisition (e.g., visual, auditory, somatosensory). It is well known that integrating complementary perceptual modalities enhances information processing and learning efficiency [21]. In remote education settings, multimodal feedback systems that support peak-referenced speech, gestures, and facial expressions excel in realizing immersive learning spaces, which helps increase learner engagement and understanding [6].

Taking into account multimodal interaction, recent studies have examined users interacting with virtual humans and highlighted the increasing significance of this in future educational environments, such as Saffaryazdi et al. [4] explored visual feedback mechanisms within virtual reality environments to improve the responsiveness and instructional capacity of digital human models. Moreover, Kotranza et al. [3] explored a virtual feedback system with peers capable of supplying multimodal feedback in real-time to users for polishing their public speaking skills.

The evolution of interactive and experience-driven models between teachers and students has become a key goal in distance education. McGraw Hill Higher Education presented a report that reflects on the education landscape in 2023; we know that the quality of education can often be ruined by poor structure; therefore, effective educators should improve the learning process by introducing active listening and adaptive feedback solutions with the aim of simplifying the active learning process for students [3]. For example, multimodal interaction technologies allow students to obtain personalized instructional guidance and real-time emotional feedback [6], creating a more engaging and effective learning experience.

This evolution dramatically accelerated the use of multimodal technology in the educational field, making possible a step towards more complete intelligent teaching tools and systems to promote interactivity and personalization in remote education opportunities.

- Interactive Teaching Tools: Multimodal technology has been integrated into intelligent learning environments like the MMISE system, which makes use of multimodal input and output modalities that consist of speech, gestures, and facial expressions to enhance the effectiveness of instruction. This system has turned out to be criti-

cally important, especially in the context of remote education during the COVID-19 pandemic [11,22].

- Improving Student Engagement: Data for students' interactions with multimodal study materials could reveal insights about comprehension and attention allocation by applying eye-tracking technologies [23]. In addition, the application of digital media entertainment technologies into remote music education has played an effective role in helping students improve their learning motivation and engagement [24,25].
- Emotion Recognition and Adaptive Feedback: By leveraging multimodal emotion recognition technology, we can identify and analyze the emotional fluctuations of students; this empowers instructors to effectively tailor their teaching strategies in real time, enhancing learning experiences [16,19].
- Remote Learning State Monitoring System: Multimodal data fusion techniques utilize speech, video, facial expressions, and gestures through deep learning models to study users' attention patterns [26,27]. Furthermore, the real-time decoding of a user's cognitive attentional states during remote learning is further made possible through the learning state monitoring system based on EEG and eye-tracking technology [28,29].
- Experimental Teaching Platforms: Platforms such as the "Remote Experimental Teaching Platform for Digital Signal Processing" enhance instructional efficacy by enabling remote program debugging and data-sharing functionalities [30]. Moreover, VR technology has been employed to develop interactive remote teaching systems, thereby improving teacher–student engagement and enhancing learning outcomes in remote education [31].
- Remote User Attention Assessment Methods: Sensor-based learning analytics approaches, including smartphone-embedded sensor technologies for the collection of data on student behavior and learning, have been integrated with embedded hardware–software systems and backend data processing frameworks to facilitate real-time dynamic assessments in remote education [31–33].
- Data Fusion and Visualization: Data fusion methodologies employ tools such as "physiological heat maps" to integrate eye-tracking and physiological data, dynamically visualizing learners' emotional and cognitive states [34]. Additionally, machine learning algorithms have been utilized to analyze students' visual attention trajectories, enabling the prediction of learning outcomes and the refinement of instructional design strategies [35].

3. Commonly Utilized Methods

3.1. Learning Indicators and Available Data

Learning analytics collects, analyzes, and reports data about learners and their environments. The goal is to better understand and improve education [36]. Since its formal introduction in 2011, this field has grown rapidly and is now used at different educational levels [37]. Its potential to improve education is clear, but its use in daily classroom activities is still slow [38].

In education, learning indicators fall into five main categories: behavior, cognition, emotion, engagement, and collaboration (Table 1).

Table 1. Learning dimensions and their descriptions.

Dimension	Description	Reference
Behavior	Learners' interaction patterns, including digital behaviors such as mouse clicks, scrolling, and text input, as well as physical actions in the learning environment	Martinez-Maldonado et al., 2018 [39]
Cognition	Learners' cognitive processes, including problem-solving ability, knowledge construction, and memory recall	Netekal et al., 2023 [40]
Emotion	Encompasses learners' emotional states such as anxiety, confidence, or frustration, which can be detected through physiological signals (e.g., EEG, EDA) or facial expression analysis	Pardo & Kloos 2011 [9]
Engagement	Measures learners' attention levels and sustained participation, often assessed through interaction behaviors or physiological data (e.g., heart rate variability)	Ido Roll & Wylie 2016 [41]
Collaboration	Describes learners' interactions in group learning, including distinctions between individual and collective attention or differences between self-regulated and collaborative learning emotions	Mu, Cui, & Huang 2020 [42]

3.2. Data Collection Methods

- Digital Interaction Data: Traditional LMS and online learning platforms primarily collect learner behavioral data through log records (clickstreams), assignment submissions, and quiz performance metrics [9].
- Physical Learning Analytics: Physical learning analytics leverages sensor technologies and the Internet of Things (IoT) to embed computational capabilities into physical environments, thereby enabling real-time interaction and data collection [43].
- Sensors and Wearable Devices: By using hard devices such as eye trackers, posture recognition cameras, and heart rate monitors, researchers can capture learners' spatial positioning, postural dynamics, and physical interactions [39]. This approach facilitates the seamless integration of learning activities across multiple environments [44], thereby extending the scope of learning analytics beyond digital interactions to encompass real-world learning contexts.
- Physiological Data: The primary objective of multimodal physiological signal research is to enhance the understanding of emotional and cognitive states by integrating multiple physiological indicators, including EEG, EDA, and electrocardiography (ECG) (Table 2).

Table 2. Physiological measurement modalities and applications.

Modality	Application	Reference
EEG	Records brain activity to analyze emotional and cognitive states	Verma & Tiwary (2014) [45]; Lin & Li (2023) [46]
EDA	Detects autonomic nervous system activity, reflecting emotional fluctuations	Horvers et al. (2021) [47]
ECG	Analyzes heart rate variability to assess emotional responses	Lin & Li (2023) [46]
Other Signals	Includes electromyography (EMG), respiratory patterns, and photoplethysmography (PPG), which enhance emotion detection accuracy in specific contexts	Verma & Tiwary (2014) [45]
Wearable Devices	Increasingly used flexible skin-attached sensors and smart devices for long-term monitoring of multiple physiological signals	Lee et al. (2019) [48]; Yang et al. (2024) [49]

3.3. Data Processing

Before using multimodal approaches, data must be integrated and analyzed. This is mainly carried out through feature-level fusion and decision-level fusion. Feature-level fusion focuses on tagging and aligning markers. It is well suited for processing experimental datasets [50,51]. Decision-level fusion is more robust and works better for multimodal analyses that involve time-based changes [52,53].

Multimodal data corpora play an important role in education research. Task-specific educational corpora collect data on interaction signals between teachers and students or among users. These signals include navigational instructions and descriptive language used during tasks [54,55].

In online learning, multimodal corpora help analyze teachers' use of visuals, gestures, and language. Studies show that gestures and visual symbols, such as arrows, help explain instructional content. This improves students' understanding and memory [56].

Developing these corpora requires well-designed experimental setups. Studies in human–computer interaction and cross-cultural behavior collect multimodal data, including gestures, intonation, and facial expressions. This expands multimodal resources for education [57,58].

4. Keyword and Correlation Analysis

To investigate the impact of multimodal interaction technology on teaching methods and instructional quality, this study selected "Multimodal Image Recognition", "Application", and "Feedback" as core keywords. A total of 25 research articles published in the past 5–10 years were manually screened to ensure relevance while maintaining broad coverage across different educational levels, specific disciplines, and target populations. Although these studies do not necessarily represent the most cutting-edge research in the field, they provide a representative sample of recent developments in the application of multimodal interaction technology in education. The selected articles were then analyzed using the VOSviewer tool to identify key trends, applications, and challenges.

4.1. Word Frequency Analysis

The analysis found 42 high-frequency terms as shown in Figure 2. The main keywords were "multimodal data", "model", "interaction", "education", and "field". These terms formed several related research clusters. They show key focus areas and reveal major research trends in multimodal interaction technology for education.

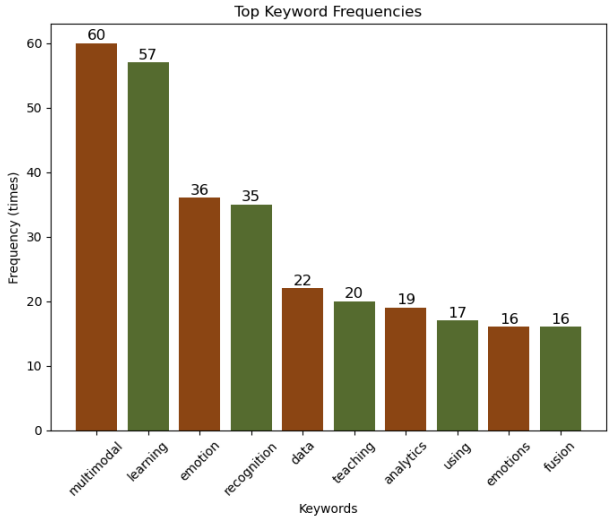

Figure 2. High-frequency research keywords.

4.2. Correlation Analysis

4.2.1. Multimodal Image Recognition

This keyword is closely associated with concepts such as accuracy, comprehension, and challenge, indicating that remote teaching environments face difficulties in ensuring the precision of image recognition and effectively understanding students' needs. Its correlation with education and students underscores the direct role of multimodal image recognition in the learning process, emphasizing the necessity of optimizing student learning experiences as shown in Figure 3. Its connection to teachers shows that feedback mechanisms help educators improve their teaching strategies.

The link between feedback and interaction highlights its key role in teaching, especially in providing personalized guidance to students. Its ties to context and interaction show the need to consider different instructional settings, such as personalized learning and adaptable curricula.

Interaction is the core of remote teaching. It requires video, audio, chat functions, and AI-assisted tools to keep students engage in virtual classrooms. Its connection to models and systems suggests that structured frameworks and technological models are essential for effective remote teaching.

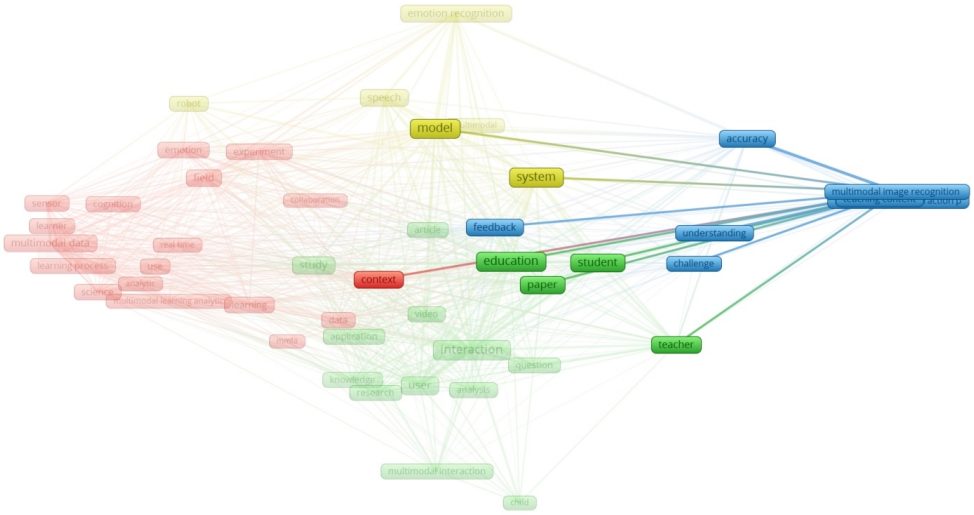

Figure 3. Correlation of "Multimodal image recognition".

4.2.2. Application

Application is a key concept that covers many fields, such as education, interaction, artificial intelligence, and learning analytics. Its link to interaction shows that interactive components play an important role in different applications, as shown in Figure 4. Its connection to users suggests that these technologies mainly support students, educators, and researchers.

A significant association with learning, cognition, and science suggests that applications in multimodal learning analytics (MMLA) contribute to improving learning experiences and facilitating personalized education. Additionally, its connection with context and video reflects the diverse operational settings of these applications, such as remote learning, which relies on video-based instruction, and mixed reality, which integrates multimodal perception technologies for immersive educational experiences.

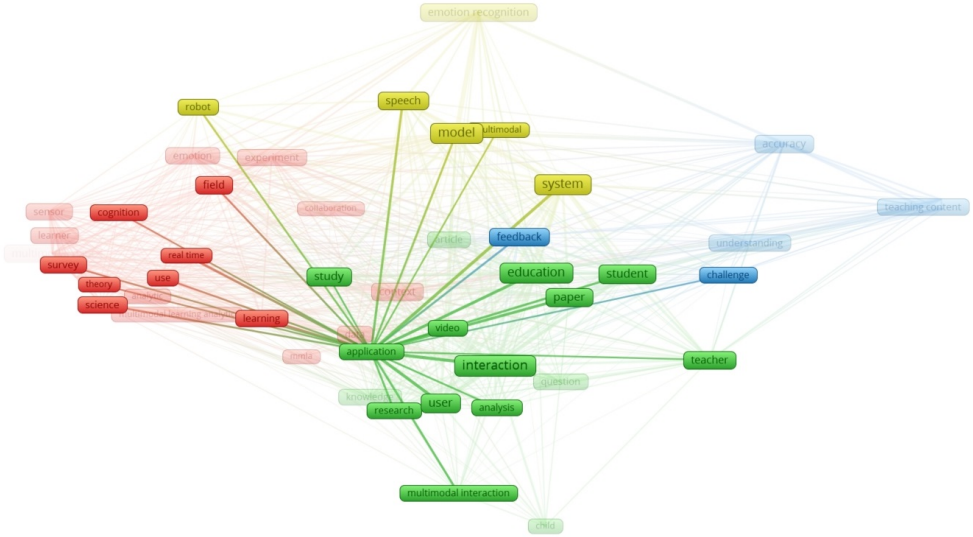

Figure 4. Correlation of "Application".

4.2.3. Feedback

The prevalence in many studies highlights its key importance in multimodal technology practice. Maintaining an efficient provision of education is an ongoing area of research, especially in student and teacher contexts where debates often revolve around understanding and challenge, emphasizing feedback in student–teacher interaction and learner adaptation. The connection between system, model, and feedback suggests that this study is examining integrating automated and AI-powered systems of feedback, including applications in speech communication and automated testing systems, as shown in Figure 5. Furthermore, the connection between learning, cognition, and feedback suggests that methods of feedback are now being created in accordance with cognitive science to better aid in learning outcomes & optimize instructional experience.

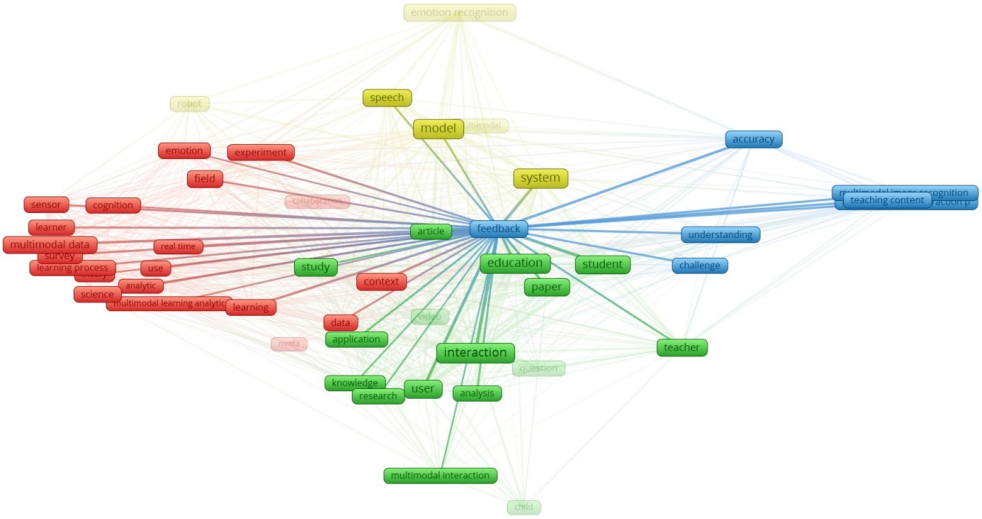

Figure 5. Correlation of "Feedback".

4.3. Multimodal Interaction Across Educational Levels

4.3.1. K-12 Education

In K-12 education, students' attention levels directly affect learning outcomes. Researchers use eye-tracking technology with speech recognition to track students' attention in real time during class. Fixation points, gaze duration, and verbal responses help educators adjust teaching strategies based on data. This improves classroom engagement

and learning effectiveness. Chen et al. [59] showed that eye-tracking technology can measure students' reading behaviors and cognitive processes, providing strong support for educational research.

4.3.2. Higher Education

In higher education and online learning, EEG and speech analysis track students' cognitive and emotional states in real time. EEG patterns and speech feedback help assess comprehension and detect emotional changes. Instructors can adjust teaching speed and offer personalized support to improve online learning. Puffay et al. [60] used deep neural networks to combine EEG and speech data, showing its potential to enhance online education.

4.3.3. Vocational Training

Vocational training focuses on practical skills. VR and AR have changed traditional training by removing time and location limits. These technologies also provide immersive remote learning.

This method increases training safety and reduces costs. It also helps trainees develop hands-on skills. In medical training, VR-based surgical simulations let trainees practice procedures and improve precision before real surgeries. Jaehyun et al. [61] studied how EEG and VR work together in emotion recognition, showing its potential in education and vocational training. Multimodal interaction technology is changing education. It improves learning efficiency and supports personalized instruction in K-12 education, higher education, and vocational training.

5. Discussion

5.1. Technical Challenges

5.1.1. Data Processing

- Synchronization Issues in Multimodal Data
  Multimodal data fusion necessitates the synchronization of data acquired from various sensors, including speech, gestures, and facial expressions. However, variations in sampling rates and timestamps among these data sources present significant synchronization challenges. For instance, speech signals typically have a higher sampling rate compared to video frames, requiring precise temporal alignment during data fusion [62,63]. Additionally, hardware-induced latency and network transmission delays further impact synchronization accuracy, complicating real-time data integration [64].
- High Computational Resource Demands and Limited Compatibility with Low-End Devices
  Processing multimodal data requires complex algorithms and deep neural networks, which demand high computational power. Real-time processing requires strong CPU and GPU performance, but many remote education devices, like personal computers and mobile devices, lack these resources. This limits the use of multimodal technology in low-resource settings. Developing lightweight and efficient algorithms can improve accessibility [65–67].
- Data Storage and Transmission Overhead
  In remote education, multimodal interaction technology produces large amounts of high-dimensional data. Managing and transmitting these datasets is a major challenge for educational platforms and cloud computing systems. Remote education depends on cloud storage and real-time streaming. High bandwidth demands can weaken system stability, especially in areas with poor network infrastructure [68,69].

5.1.2. User Experience

- Teachers' Adaptation to Multimodal Technology
  Multimodal technology has changed traditional teaching methods, and thus teachers have to learn new tools and platforms. Some instructors struggle to leverage these tools to their full potential because they have not been trained in them. This inhibits their ability to fully benefit from multimodal systems [70]. Hardware failures and computer glitches lead to added workloads for instructors and compromised teaching effectiveness [71,72].
- Students' Acceptance of Personalized Learning Pathways
  Multimodal technology facilitates the development of personalized learning pathways by adapting instructional content to students' individual learning styles and progress. However, students' acceptance of such personalized learning approaches varies. While some learners prefer self-directed learning and customized educational content, others may be more inclined toward traditional teacher-led instruction [73]. Furthermore, designing personalized learning pathways necessitates a balance between self-regulated learning and teacher guidance to ensure an optimal and structured learning experience [64].
- Complexity of Human–Computer Interaction
  Multimodal systems support different input modes beyond speech, including touch interfaces, gestures, and facial recognition. Users must switch between these modes to interact with the system. Some users face problems with usability. Common issues include complex interface design, system delays after input, and low recognition accuracy [74]. High error rates also reduce user experience. Inaccurate speech recognition can cause unintended actions, and poor gesture recognition can disrupt classroom teaching. A major challenge in multimodal instructional systems is maintaining a good user experience while keeping the system easy to use [27].

5.1.3. Privacy and Ethical Issues

- Privacy Protection
  Physiological student data in remote teaching multimodal environments are employed extensively to determine learning status and adjust teaching practices. However, the sensitivity of this type of data is extremely high. Today, most remote learning systems have incomplete security measures to protect physiological data, making them vulnerable to unauthorized use, illegal access, and potential breaches. The application of data protection laws, such as GDPR, in education is unclear. This creates legal risks in data storage, processing, and use. A major challenge is protecting data privacy while making full use of multimodal data to improve education [75].
- Algorithmic Fairness
  Multimodal learning systems are prone to algorithmic biases, which can affect fairness and inclusivity. AI models are trained on historical datasets that may contain cultural, gender, or linguistic biases. These biases can lead to unfair feedback or unequal learning experiences for students from different backgrounds. Some speech recognition systems have difficulty processing non-native speakers' inputs, which can negatively impact their learning. Adaptive learning systems may also create inequalities. Personalized recommendation algorithms might assign different difficulty levels based on existing gaps, reinforcing disparities [76]. Future research should focus on improving fairness in multimodal AI algorithms to provide equal educational support for all students.
- Security Concerns

An increase in remote education will mean that cloud-based systems used to store and process data are vulnerable to cyber attacks, data theft, or identity theft. The security and privacy threats of unauthorized access to student behavioral data, learning records, physiological information, etc., are serious. Distributed ledger technology (DLT). The existing encryption techniques and access control mechanisms should be further optimized to meet the needs and requirements of the increasing rate of attack on new remote education systems. Technologies such as federated learning, which allow for decentralized data usage, help reduce the risk of data falling, while advanced blockchain-based identity authentication provides stronger data security and integrity [77]. This foundation of sustainability and resilience depends on which multimodal curricula are designed, and this is where the data security framework must be laid.

5.2. Limitations of Current Research

Although there have been great advances in multimodal interaction technology for remote education, there are still several limitations in existing studies, as follows:

- Diversity and Representativeness of Data
  Existing multimodal learning datasets are mostly acquired in controlled experimental environments, while there is no adequately large-scale data available from authentic classroom settings. This limitation restricts the external validity and generalizability of research findings. Immadisetty et al. [64] emphasized the importance of capturing diverse datasets in real-world educational settings to advance the relevance of multimodal learning studies.
- Exploiting the Gap Between Experimental Settings and Practical Utilization
  Although randomized experiments are used to measure a wide variety of interventions to improve student learning, these interventions are often implemented in highly controlled laboratory environments that do not reflect the complexity of real educational settings. This divergence can lead to experimental results that do not translate well to the field. Liu et al. [78] emphasized the need to validate multimodal interaction systems in real educational scenarios to ensure their real-world effectiveness and scalability.
- Assessment of Long-Term Learning Outcomes
  Much of the existing research focuses on short-term experimental studies, and few studies have conducted rigorous longitudinal investigations covering the sustained effect of multimodal interaction technology. Future studies should include long-term follow-up investigations of its impact on the ongoing cognitive and behavioral development of learners.

Table 3 compares the limitations of existing studies.

Most studies were conducted in labs or controlled environments. Only a few were conducted in real classrooms or remote learning settings. Many studies used speech or gesture as the main interaction methods. Some used sensors like EEG or eye tracking. These systems often focused on college students. Fewer studies looked at younger learners or people with special needs.

Many systems did not give real-time feedback. Some had small sample sizes. Others used short-term tests. Most studies did not include long-term data or results from large user groups.

Because most research was carried out in ideal conditions, the findings might not match those from real-world classrooms. The systems might work differently when students are at home or using their own devices. Some studies did not talk about cost or technical limitations.

Table 3. Comparative summary of selected studies on multimodal interaction in remote education.

Study	Modality Types Used	Education Level	Target Group	Environment	Limitations
Immadisetty et al., 2023 [64]	Posture and gesture recognition, facial analysis, eye-tracking, verbal recognition	Higher Education	General students	Controlled laboratory setting	Limited real-world classroom deployment
Faridan et al., 2023 [79]	Mixed reality gestural guidance	Higher Education	Physiotherapy students	Simulated classroom environment	Small sample size
Zhang et al., 2024 [80]	LLM-empowered agents simulating teacher-student interactions	Higher Education	General students	Virtual classroom simulation	Lack of real-world validation
Hao et al., 2021 [81]	Pre-trained language models for dialogic instruction detection	Higher Education	Online learners	Online educational platform	Focused on text-based interactions only
Li et al., 2020 [82]	Machine learning for identifying at-risk students	K-12 Education	K-12 students	Multimodal online environments	Data imbalance, limited offline factors

5.3. Future Work

- Tailoring Adaptive Learning Journeys
  Using emotion recognition and behavioral analysis, the existing multimodal education systems aim to provide real-time adaptive feedback, but more studies are needed to fine-tune personalized learning paths with better use of multimodal data. For instance, Liu et al. [78] developed a personalized multimodal feedback generation network that integrates multiple modal inputs to produce customized feedback on student assignments, ultimately improving learning efficiency.
- Flexibility when Deployed in Low-Resource Contexts
  Multimodal interaction technology frequently requires high-performance computing resources, making its deployment in resource-constrained regions difficult. Immadisetty et al. [64] highlighted the vital need to develop multimodal education systems tailored to low-bandwidth, low-computation environments, thus fostering educational equity and technological access.
- International Adaptability
  Additionally, existing multimodal interaction systems are typically designed within specific cultural and linguistic contexts, which may restrict cross-cultural relevance and scalability (e.g., in multicultural educational environments). Future research should focus on developing multimodal systems that are adaptable to diverse cultural contexts, ensuring broader usability and acceptance across various educational institutions.

6. Conclusions

This research investigates the impact of multimodal interaction technology on remote education, focusing on its benefits for enriching learning experiences, tailoring instruction, and integrating affective computing. The results suggest that the fusion of multimodal data can significantly enhance interaction and teaching efficacy in remote learning settings.

Nonetheless, challenges persist, including issues with data synchronization, the demand for computational resources, user adaptability, and concerns regarding privacy and security. One significant challenge is synchronizing multimodal data from various sources, as differences in sampling rates and temporal misalignment can compromise real-time accuracy. In addition, high computational demands often hinder the practical deployment of these systems on low-resource devices commonly used in educational settings. Moreover, the complexity of human–computer interaction and the varying acceptance of personalized learning pathways can pose usability issues. Finally, privacy protection and algorithmic fairness remain critical concerns, as the extensive use of multimodal data increases the risk of misuse and discrimination.

While the potential of this technology is substantial, existing research is constrained. A majority of studies are conducted in controlled laboratory environments, which limits the availability of large-scale, real-world classroom data and diminishes their relevance in practical teaching contexts. Furthermore, there is a notable deficiency in assessing the long-term effects of multimodal interaction technology, highlighting the need for additional research into its enduring influence on students' cognitive and behavioral growth.

Future investigations should prioritize the optimization of personalized learning pathways, the enhancement of adaptability in resource-limited settings, and the improvement of cross-cultural applicability to promote the widespread adoption of this technology in remote education. As advancements in artificial intelligence and sensor technologies progress, multimodal interaction technology is poised to transform remote education into a more intelligent and personalized learning experience.

Author Contributions: Writing—original draft preparation, Y.X., L.Y., M.Z., S.C. and J.L.; writing—review and editing, Y.X., L.Y., M.Z., S.C. and J.L.; visualization, Y.X.; supervision, S.C. and J.L.; project administration, S.C. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by Kobe University CMDS Joint Project Promoting DX Inside & Outside the University. Grant Number PJ2024-03.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Obrenovic, Z.; Starcevic, D. Modeling multimodal human-computer interaction. *Computer* **2004**, *37*, 65–72. [CrossRef]
2. Saffaryazdi, N.; Goonesekera, Y.; Saffaryazdi, N.; Hailemariam, N.D.; Temesgen, E.G.; Nanayakkara, S.; Broadbent, E.; Billinghurst, M. Emotion Recognition in Conversations Using Brain and Physiological Signals. In Proceedings of the IUI'22: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, 22–25 March 2022; pp. 229–242. [CrossRef]
3. Chollet, M.; Wörtwein, T.; Morency, L.P.; Shapiro, A.; Scherer, S. Exploring feedback strategies to improve public speaking: An interactive virtual audience framework. In Proceedings of the UbiComp'15: 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Osaka, Japan, 7–11 September 2015; pp. 1143–1154. [CrossRef]
4. Geiger, A.; Bewersdorf, I.; Brandenburg, E.; Stark, R. Visual Feedback for Grasping in Virtual Reality Environments for an Interface to Instruct Digital Human Models. In Proceedings of the Advances in Usability and User Experience, Barcelona, Spain, 19–20 September 2018; Ahram, T., Falcão, C., Eds.; Springer: Cham, Swirzerlands, 2018; pp. 228–239.
5. Kotranza, A.; Lok, B.; Pugh, C.M.; Lind, D.S. Virtual Humans That Touch Back: Enhancing Nonverbal Communication with Virtual Humans through Bidirectional Touch. In Proceedings of the 2009 IEEE Virtual Reality Conference, Lafayette, LA, USA, 14–18 March 2009; pp. 175–178. [CrossRef]
6. Wigdor, D.; Wixon, D. *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*; Elsevier: Amsterdam, The Netherlands, 2011.

7. Schwendimann, B.A.; Rodríguez-Triana, M.J.; Vozniuk, A.; Prieto, L.P.; Boroujeni, M.S.; Holzer, A.; Gillet, D.; Dillenbourg, P. Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research. *IEEE Trans. Learn. Technol.* **2017**, *10*, 30–41. [CrossRef]
8. Di Mitri, D.; Schneider, J.; Specht, M.; Drachsler, H. From signals to knowledge: A conceptual model for multimodal learning analytics. *J. Comput. Assist. Learn.* **2018**, *34*, 338–349. [CrossRef]
9. Pardo, A.; Kloos, C.D. Stepping out of the box: Towards analytics outside the learning management system. In Proceedings of the LAK'11: 1st International Conference on Learning Analytics and Knowledge, New York, NY, USA, 27 February–1 March 2011; pp. 163–167. [CrossRef]
10. Di Mitri, D. Digital Learning Projection. In *Artificial Intelligence in Education*; André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B., Eds.; Springer: Cham, Swirzerlands, 2017; pp. 609–612.
11. Jia, J.; Yunfan, H.; Huixiao, L. A Multimodal Human-Computer Interaction System and Its Application in Smart Learning Environments; Springer: Cham, Swirzerlands, 2020; pp. 3–14. [CrossRef]
12. Lazaro, M.J.; Kim, S.; Lee, J.; Chun, J.; Kim, G.; Yang, E.; Bilyalova, A.; Yun, M.H. A Review of Multimodal Interaction in Intelligent Systems. In *Human-Computer Interaction. Theory, Methods and Tools*; Kurosu, M., Ed.; Springer: Cham, Swirzerlands, 2021; pp. 206–219.
13. Jiang, Y.; Li, W.; Hossain, M.S.; Chen, M.; Alelaiwi, A.; Al-Hammadi, M. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Inf. Fusion* **2020**, *53*, 209–221. [CrossRef]
14. Nguyen, R.; Gouin-Vallerand, C.; Amiri, M. Hand interaction designs in mixed and augmented reality head mounted display: A scoping review and classification. *Front. Virtual Real.* **2023**, *4*, 1171230. [CrossRef]
15. Tu, Y.; Luo, J. *Accessibility Research on Multimodal Interaction for the Elderly*; ACM: New York, NY, USA, 2024; pp. 384–398. [CrossRef]
16. Ramaswamy, M.P.A.; Palaniswamy, S. Multimodal emotion recognition: A comprehensive review, trends, and challenges. *WIREs Data Min. Knowl. Discov.* **2024**, *14*, e1563. [CrossRef]
17. Koromilas, P.; Giannakopoulos, T. Deep Multimodal Emotion Recognition on Human Speech: A Review. *Appl. Sci.* **2021**, *11*, 7962. [CrossRef]
18. Wang, T.; Zheng, P.; Li, S.; Wang, L. Multimodal Human–Robot Interaction for Human-Centric Smart Manufacturing: A Survey. *Adv. Intell. Syst.* **2024**, *6*, 2300359. [CrossRef]
19. Pan, B.; Hirota, K.; Jia, Z.; Dai, Y. A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing* **2023**, *561*, 126866. [CrossRef]
20. Zhu, X.; Guo, C.; Feng, H.; Huang, Y.; Feng, Y.; Wang, X.; Wang, R. A Review of Key Technologies for Emotion Analysis Using Multimodal Information. *Cogn. Comput.* **2024**, *16*, 1504–1530. [CrossRef]
21. Calvo, R.; D'Mello, S.; Gratch, J.; Kappas, A. *The Oxford Handbook of Affective Computing*; Oxford Academic: Oxford, UK, 2014.
22. Cornide-Reyes, H.; Riquelme, F.; Monsalves, D.; Noël, R.; Cechinel, C.; Villarroel, R.; Ponce, F.; Muñoz, R. A Multimodal Real-Time Feedback Platform Based on Spoken Interactions for Remote Active Learning Support. *Sensors* **2020**, *20*, 6337. [CrossRef]
23. Gatcho, A.; Manuel, J.P.; Sarasua, R. Eye tracking research on readers' interactions with multimodal texts: A mini-review. *Front. Commun.* **2024**, *9*, 1482105. [CrossRef]
24. Wang, J. Application of digital media entertainment technology based on soft computing in immersive experience of remote piano teaching. *Entertain. Comput.* **2025**, *52*, 100822. [CrossRef]
25. Choo, Y.B.; Saidalvi, A.; Abdullah, T. Use of Multimodality in Remote Drama Performance among Pre-Service Teachers during the Covid-19 Pandemic. *Int. J. Acad. Res. Progress. Educ. Dev.* **2022**, *11*, 743–760. [CrossRef]
26. Ranjan, R.; Patel, V.M.; Chellappa, R. HyperFace: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 121–135. [CrossRef]
27. Cruciata, G.; Lo Presti, L.; Cascia, M.L. On the Use of Deep Reinforcement Learning for Visual Tracking: A Survey. *IEEE Access* **2021**, *9*, 120880–120900. [CrossRef]
28. Miran, S.; Akram, S.; Sheikhattar, A.; Simon, J.Z.; Zhang, T.; Babadi, B. Real-time tracking of selective auditory attention from M/EEG: A bayesian filtering approach. *Front. Neurosci.* **2018**, *12*, 262. [CrossRef]
29. Tanaka, N.; Watanabe, K.; Ishimaru, S.; Dengel, A.; Ata, S.; Fujimoto, M. Concentration Estimation in Online Video Lecture Using Multimodal Sensors. In Proceedings of the Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing, Melbourne, VIC, Australia, 5–9 October 2024. [CrossRef]
30. Shang, Q.; Zheng, G.; Li, Y. Mobile Learning Based on Remote Experimental Teaching Platform. In Proceedings of the ICEMT'19: 3rd International Conference on Education and Multimedia Technology, Nagoya, Japan, 22–25 July 2019; p. 307–310. [CrossRef]
31. Ge, T.; Darcy, O. Study on the Design of Interactive Distance Multimedia Teaching System based on VR Technology. *Int. J. Contin. Eng. Educ.-Life-Long Learn.* **2021**, *31*, 1. [CrossRef]

32. Engelbrecht, J.M.; Michler, A.; Schwarzbach, P.; Michler, O. Bring Your Own Device-Enabling Student-Centric Learning in Engineering Education by Incorporating Smartphone-Based Data Acquisition. In Proceedings of the International Conference on Interactive Collaborative Learning, Madrid, Spain, 26–29 September 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 373–383.
33. Javeed, M.; Mudawi, N.A.; Alazeb, A.; Almakdi, S.; Alotaibi, S.S.; Chelloug, S.; Jalal, A. Intelligent ADL Recognition via IoT-Based Multimodal Deep Learning Framework. *Sensors* **2023**, *23*, 7927. [CrossRef]
34. Courtemanche, F.; Léger, P.M.; Dufresne, A.; Fredette, M.; Labonté-LeMoyne, É.; Sénécal, S. Physiological heatmaps: A tool for visualizing users' emotional reactions. *Multimed. Tools Appl.* **2018**, *77*, 11547–11574. [CrossRef]
35. Chettaoui, N.; Atia, A.; Bouhlel, M.S. Student performance prediction with eye-gaze data in embodied educational context. *Educ. Inf. Technol.* **2023**, *28*, 833–855.
36. Siemens, G.; Long, P. Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Rev.* **2011**, *5*, 30–32. [CrossRef]
37. Gašević, D.; Dawson, S.; Pardo, A. *How Do We Start? State Directions of Learning Analytics Adoption*; ICDE—International Council For Open And Distance Education: Oslo, Norway, 2016.
38. Gašević, D.; Dawson, S.; Siemens, G. Let's not forget: Learning analytics are about learning. *TechTrends* **2015**, *59*, 64–71.
39. Martinez-Maldonado, R.; Echeverria, V.; Santos, O.C.; Santos, A.D.P.D.; Yacef, K. Physical learning analytics: A multimodal perspective. In Proceedings of the LAK'18: 8th International Conference on Learning Analytics and Knowledge, Sydney, NSW, Australia, 7–9 March 2018; pp. 375–379. [CrossRef]
40. Netekal, M.; Hegade, P.; Shettar, A. Knowledge Structuring and Construction in Problem Based Learning. *J. Eng. Educ. Transform.* **2023**, *36*, 186–193. [CrossRef]
41. Roll, I.; Wylie, R. Evolution and revolution in artificial intelligence in education. *Int. J. Artif. Intell. Educ.* **2016**, *26*, 582–599.
42. Mu, S.; Cui, M.; Huang, X. Multimodal Data Fusion in Learning Analytics: A Systematic Review. *Sensors* **2020**, *20*, 6856. [CrossRef]
43. Ebling, M.R. Pervasive Computing and the Internet of Things. *IEEE Pervasive Comput.* **2016**, *15*, 2–4. [CrossRef]
44. Kurti, A.; Spikol, D.; Milrad, M.; Svensson, M.; Pettersson, O. Exploring How Pervasive Computing Can Support Situated Learning. In Proceedings of the Pervasive Learning 2007, Toronto, ON, Canada, 13 May 2007.
45. Verma, G.K.; Tiwary, U.S. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage* **2014**, *102*, 162–172. [CrossRef]
46. Li, L.; Gui, X.; Huang, G.; Zhang, L.; Wan, F.; Han, X.; Wang, J.; Ni, D.; Liang, Z.; Zhang, Z. Decoded EEG neurofeedback-guided cognitive reappraisal training for emotion regulation. *Cogn. Neurodyn.* **2024**, *18*, 2659–2673.
47. Horvers, A.; Tombeng, N.; Bosse, T.; Lazonder, A.; Molenaar, I. Detecting Emotions through Electrodermal Activity in Learning Contexts: A Systematic Review. *Sensors* **2021**, *21*, 7869. [CrossRef]
48. Lee, M.; Cho, Y.; Lee, Y.; Pae, D.; Lim, M.; Kang, T.K. PPG and EMG Based Emotion Recognition using Convolutional Neural Network, In Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics, Prague, Czech Republic, 29–31 July 2019; pp. 595–600. [CrossRef]
49. Yang, G.; Kang, Y.; Charlton, P.H.; Kyriacou, P.; Kim, K.K.; Li, L.; Park, C. Energy-Efficient PPG-Based Respiratory Rate Estimation Using Spiking Neural Networks. *Sensors* **2024**, *24*, 3980. [CrossRef] [PubMed]
50. Heideklang, R.; Shokouhi, P. Fusion of multi-sensory NDT data for reliable detection of surface cracks: Signal-level vs. decision-level. *AIP Conf. Proc.* **2016**, *1706*, 180004. [CrossRef]
51. Uribe, Y.F.; Alvarez-Uribe, K.C.; Peluffo-Ordóñez, D.H.; Becerra, M.A. Physiological Signals Fusion Oriented to Diagnosis—A Review. In *Communications in Computer and Information Science*; Springer: Cham, Switzerlands, 2018; pp. 1–15. [CrossRef]
52. Sad, G.D.; Terissi, L.D.; Gómez, J. Decision Level Fusion for Audio-Visual Speech Recognition in Noisy Conditions. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*; Springer: Cham, Switzerlands, 2016; pp. 360–367. [CrossRef]
53. Rabbani, M.H.R.; Islam, S.M.R. Multimodal Decision Fusion of EEG and fNIRS Signals. In Proceedings of the 2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), Dhaka, Bangladesh, 18–20 November 2021; pp. 1–6. [CrossRef]
54. Xu, P.; Jiang, H. Review of the Application of Multimodal Biological Data in Education Analysis. *IOP Conf. Ser. Earth Environ. Sci.* **2018**, *170*, 022175. [CrossRef]
55. Chango, W.; Lara, J.; Cerezo, R.; Romero, C. A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2022**, *12*, e1458. [CrossRef]
56. Blikstein, P.; Worsley, M. Multimodal Learning Analytics and Education Data Mining: Using computational technologies to measure complex learning tasks. *J. Learn. Anal.* **2016**, *3*, 220–238. [CrossRef]
57. Doumanis, I.; Economou, D.; Sim, G.; Porter, S. The impact of multimodal collaborative virtual environments on learning: A gamified online debate. *Comput. Educ.* **2019**, *130*, 121–138. [CrossRef]
58. Perveen, A. Facilitating Multiple Intelligences Through Multimodal Learning Analytics. *Turk. Online J. Distance Educ.* **2018**, *19*, 18–30. [CrossRef]

59. Chen, Y.; Guo, Q.; Qiao, C.; Wang, J. A systematic review of the application of eye-tracking technology in reading in science studies. *Res. Sci. Technol. Educ.* **2023**, 1–25. [CrossRef]
60. Puffay, C.; Accou, B.; Bollens, L.; Monesi, M.J.; Vanthornhout, J.; Hamme, H.V.; Francart, T. Relating EEG to continuous speech using deep neural networks: A review. *arXiv* **2023**, arXiv:2302.01736.
61. Nam, J.; Chung, H.; ah Seong, Y.; Lee, H. A New Terrain in HCI: Emotion Recognition Interface using Biometric Data for an Immersive VR Experience. *arXiv* **2019**, arXiv:1912.01177.
62. Basystiuk, O.; Rybchak, Z.; Zavushchak, I.; Marikutsa, U. Evaluation of multimodal data synchronization tools. *Comput. Des. Syst. Theory Pract.* **2024**, *6*, 104–111. [CrossRef]
63. Malawski, F.; Kapela, K.; Krupa, M. Synchronization of External Inertial Sensors and Built-in Camera on Mobile Devices. In Proceedings of the 2023 IEEE Symposium Series on Computational Intelligence (SSCI), Mexico City, Mexico, 5–8 December 2023; pp. 772–777. [CrossRef]
64. Immadisetty, P.; Rajesh, P.; Gupta, A.; MR, D.A.; Subramanya, D.K. Multimodality in online education: A comparative study. *Multimed. Tools Appl.* **2025**, 1–34. [CrossRef]
65. Lyu, Y.; Zheng, X.; Kim, D.; Wang, L. OmniBind: Teach to Build Unequal-Scale Modality Interaction for Omni-Bind of All. *arXiv* **2024**, arXiv:2405.16108.
66. Cui, K.; Zhao, M.; He, M.; Liu, D.; Zhao, Q.; Hu, B. MultimodalSleepNet: A Lightweight Neural Network Model for Sleep Staging based on Multimodal Physiological Signals. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 3–6 December 2024; pp. 5264–5271. [CrossRef]
67. Guo, Z.; Ma, H.; Li, A. A lightweight finger multimodal recognition model based on detail optimization and perceptual compensation embedding. *Comput. Stand. Interfaces* **2024**, *92*, 103937. [CrossRef]
68. Ranhel, J.; Vilela, C. Guidelines for creating man-machine multimodal interfaces. *arXiv* **2020**, arXiv:1901.10408.
69. Agrawal, D.; Kalpana, C.; Lachhani, M.; Salgaonkar, K.A.; Patil, Y. Role of Cloud Computing in Education. *REST J. Data Anal. Artif. Intell.* **2023**, *2*, 38–42. [CrossRef]
70. Qushem, U.B.; Christopoulos, A.; Oyelere, S.; Ogata, H.; Laakso, M. Multimodal Technologies in Precision Education: Providing New Opportunities or Adding More Challenges? *Educ. Sci.* **2021**, *11*, 338. [CrossRef]
71. Wu, Y.; Sun, Y.; Sundar, S.S. What Do You Get from Turning on Your Video? Effects of Videoconferencing Affordances on Remote Class Experience During COVID-19. *Proc. ACM Hum.-Comput. Interact.* **2022**, *6*, 3555773. [CrossRef]
72. Manamela, L.E.; Sumbane, G.O.; Mutshatshi, T.E.; Ngoatle, C.; Rasweswe, M.M. Multimodal teaching and learning challenges: Perspectives of undergraduate learner nurses at a higher education institution in South Africa. *Afr. J. Health Prof. Educ.* **2024**, *16*, e1299. [CrossRef]
73. Khor, E.; Tan, L.P.; Chan, S.H.L. Systematic Review on the Application of Multimodal Learning Analytics to Personalize Students' Learning. *AsTEN J. Teach. Educ.* **2024**, 1–16. [CrossRef]
74. Darin, T.G.R.; Andrade, R.; Sánchez, J. Usability evaluation of multimodal interactive virtual environments for learners who are blind: An empirical investigation. *Int. J. Hum. Comput. Stud.* **2021**, *158*, 102732. [CrossRef]
75. Liu, D.K. Predicting Stress in Remote Learning via Advanced Deep Learning Technologies. *arXiv* **2021**, arXiv:2109.11076.
76. Peng, X.; Wei, Y.; Deng, A.; Wang, D.; Hu, D. Balanced multimodal learning via on-the-fly gradient modulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8238–8247.
77. Liu, Y.; Li, K.; Huang, Z.; Li, B.; Wang, G.; Cai, W. EduChain: A blockchain-based education data management system. In Proceedings of the Blockchain Technology and Application: Third CCF China Blockchain Conference, CBCC 2020, Jinan, China, 18–20 December 2020; Revised Selected Papers 3. Springer: Berlin/Heidelberg, Germany, 2021; pp. 66–81.
78. Liu, H.; Liu, Z.; Wu, Z.; Tang, J. Personalized Multimodal Feedback Generation in Education. *arXiv* **2020**, arXiv:2011.00192.
79. Faridan, M.; Kumari, B.; Suzuki, R. ChameleonControl: Teleoperating Real Human Surrogates through Mixed Reality Gestural Guidance for Remote Hands-on Classrooms. In Proceedings of the CHI'23: 2023 CHI Conference on Human Factors in Computing Systems, ACM, Hamburg, Germany, 23–28 April 2023; pp. 1–13. [CrossRef]
80. Zhang, Z.; Zhang-Li, D.; Yu, J.; Gong, L.; Zhou, J.; Hao, Z.; Jiang, J.; Cao, J.; Liu, H.; Liu, Z.; et al. Simulating Classroom Education with LLM-Empowered Agents. *arXiv* **2024**, arXiv:2406.19226.
81. Hao, Y.; Li, H.; Ding, W.; Wu, Z.; Tang, J.; Luckin, R.; Liu, Z. Multi-Task Learning based Online Dialogic Instruction Detection with Pre-trained Language Models. *arXiv* **2021**, arXiv:2107.07119.
82. Li, H.; Ding, W.; Liu, Z. Identifying At-Risk K-12 Students in Multimodal Online Environments: A Machine Learning Approach. *arXiv* **2020**, arXiv:2003.09670.