Computers and Education: Artificial Intelligence 5 (2023) 100187

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Designing an artificial intelligence tool to understand student engagement based on teacher's behaviours and movements in video conferencing

Navdeep Verma [a,*], Dr Seyum Getenet [b], Dr Christopher Dann [b], Thanveer Shaik [c]

[a] University of Southern Queensland, Springfield Education City, 37 Sinnathamby Blvd, Springfield Central, QLD, 4300, Australia
[b] University of Southern Queensland, School of Education, Australia
[c] University of Southern Queensland, School of Maths, Physics & Computing, Australia

ARTICLE INFO

Keywords:
Artificial intelligence
Video conferencing
Teachers' behaviours
Teachers' movements

ABSTRACT

Video conferencing is an effective tool that promotes interaction and collaboration, increasing student engagement in online learning. This study is the second phase of design-based research to create a tool to generate a report of engaging teaching videos using deep learning as an artificial intelligence (AI) methodology. In this second phase, the authors have applied the characteristics and indicators of engaging teaching videos identified in the first phase, reported in another study, to develop an Artificial Intelligence enabled tool. Twenty-five recorded lecture videos presented to higher education students were annotated based on the indicators and characteristics of engaging teaching videos. An AI expert has assisted the authors in creating the Artificial Intelligence-enabled tool from the reports generated by this manual annotation. With the assistance of this tool, the engagement enhancing teachers' behaviours and movements can be identified from recorded lecture videos, and a report can be generated on engaging teaching videos. For the classification task of video analysis, the deep learning model is adopted in this research. The model is trained with manually annotated videos and determines class imbalance issues and misleading metrics. The model was further improved by adopting the oversampling technique. The second version of the tool achieved promising outputs with average precision, recall, f1-score, and balanced accuracy of 68, 75, 73, and 79%, respectively, in classifying the annotated videos at the indicator level. The tool can assist the education institutes in creating moderation in the lecture delivery and whether the teachers are utilising the technology effectively. Additionally, this can help teachers recognise the presence or absence of engagement-enhancing behaviours and movements during their video conferencing sessions.

1. Introduction

Over the past few decades, the demand for online learning has significantly increased in higher education, providing students and teachers with more flexible access to educational opportunities. In 2020, due to the global issue of COVID-19, higher education institutes worldwide were compelled to switch their learning mode to online learning (Cucinotta & Vanelli, 2020). This sudden change caused many challenges to teachers and students as they were not prepared for online learning. The research indicates that student engagement in online learning is a significant challenge, and engaging them in online learning is more complex than face-to-face learning (Cesari et al., 2021; Gillett-Swan, 2017; Hew, 2016).

In online learning, technology is crucial in delivering education and enhancing interaction (Singh & Thurman, 2019). Video conferencing is the most effective technology that assists teachers in offering collaboration and increasing student engagement (Kumar et al., 2015); it is an effective instrument for teaching and communication in online learning (Al-Samarraie, 2019). Several video conferencing platforms are available, such as Zoom, WebEx, Microsoft Teams, GoTo Meeting and Skype for business (Döring et al., 2022). These platforms allow teachers and learners to communicate in real-time via live audio and video (Lieux et al., 2021). This real-time connectivity enhances human connection, permitting educators and students to establish their presence in online learning (Burke et al., 2022). By utilising video conferencing effectively, teachers can provide immediate feedback to students to bridge the psychological and communication distance between them and students (Torrato et al., 2021). Video conferencing also allows wireless screen

* Corresponding author.
E-mail addresses: Navdeep.Verma@usq.edu.au (N. Verma), Seyum.Getenet@usq.edu.au (D.S. Getenet), Christopher.Dann@usq.edu.au (D.C. Dann), Thanveer.Shaik@usq.edu.au (T. Shaik).

https://doi.org/10.1016/j.caeai.2023.100187
Received 25 June 2023; Received in revised form 16 November 2023; Accepted 18 November 2023
Available online 23 November 2023

sharing, whiteboard sharing, interactive chat rooms, opinion polls, and discussion platforms.

In their research, Wang et al. (2018) stated that students' engagement is highly needed to successfully utilise video Conferencing tools and conduct online learning. Teachers can use engagement enhancing behaviour and movements to engage students in video conferencing. These behaviours consist of (a combination of) autonomy support, a structure to enhance students' sense of competence, and relatedness support (Aelterman et al., 2019; De Meester et al., 2020). Some of these motivating teachers' behaviour are encouraging active participation, establishing teacher presence, demonstrating empathy, and establishing clear expectations (Authors, 2023). Furthermore, teachers' positive body movements include non-verbal cues, such as facial expressions, gestures, and eye movements (Authors, 2023).

Previous studies have established that these behaviours and movements enhance teachers' presence, increasing student engagement (Cents-Boonstra et al., 2021; Dewan et al., 2019). Therefore, there is a need for an instrument that can measure engaging teaching videos according to teachers' movements and behaviours. The authors have identified these behaviours and movements through a systematic literature review in phase 1, reported in another study (Authors, 2023). Identifying these engaging characteristics and indicators from recorded lecture videos requires a human, and this manual identification and analysis is very time-consuming and expensive (Beaver & Mueen, 2022), and it can also lead to human bias.

To avoid human bias and to measure engaging teaching videos much faster, the authors employed a designed-based research (DBR) approach to create an AI-enabled tool that generates a report for indicators and characteristics of engaging teaching videos. The identified indicators and characteristics in phase 1 of DBR research (Authors, 2023) were applied to 25 Zoom-recorded lecture videos through a manual annotation process. With manual annotation outputs (reports), the AI expert trained the AI-enabled tool. This tool will assist teachers in enhancing the overall quality of online learning. Educational institutes can also utilise this tool to create moderation in their lecture delivery and improve online learning procedures.

The study explored the following research questions.

(1). To what extent can an AI-enabled tool be designed to generate a report for indicators and characteristics of engaging teaching videos based on teachers' behaviours and movements?
(2). How will the AI-enabled tool improve teaching and learning practices in higher education?

2. Literature review

2.1. Online learning in higher education

In today's technology-driven environment, online learning has gained increasing attention. A growing number of students are choosing online learning, leading to online learning in higher education becoming a primary mode of delivery (Salas-Pilco et al., 2022). Further, COVID-19 has created a radical change in education, and higher education worldwide is going through a dramatic transformation in digital technologies (Dwivedi et al., 2020; Krishnamurthy, 2020). Even though the use of technology began several years ago, the abrupt change caused by this pandemic has required higher education institutions worldwide to transition to online learning rapidly. This evolution includes integrating and using technological resources available to teachers and researchers (García-Morales et al., 2021).

2.2. Video conferencing and student engagement

In online learning, technology develops a connection between teachers and students and assists in developing the abilities required for online learning. Due to its ability to provide real-time communication through audio and video calls (Lieux et al., 2021), video conferencing is now an essential tool in online learning (Roth et al., 2020). Several video conferencing tools are available for teachers to digitalise their teaching and learning processes (Mishra et al., 2020). These tools are GoTo Meeting, Microsft Teams, WebEx, Zoom and Skype (Döring et al., 2022). Although these tools provide various functions to improve online teaching and learning, student engagement is crucial for enhancing teaching quality and video conferencing.

Student engagement is essential for student achievement, academic motivation, performance, and satisfaction (Hu & Li, 2017, pp. 39–43; Kahu et al., 2019). For deep-level learning and retention, engaging students in learning is necessary (Cents-Boonstra et al., 2021; Hu & Li, 2017, pp. 39–43; Kuh et al., 2008). However, the lack of student engagement is a significant issue in online learning (Dembereldorj, 2021), which results in non-completion, withdrawal, and unsatisfactory learning experiences (Bergdahl, 2022).

Video conferencing can assist in engaging students in online learning by offering features such as whiteboard sharing, screen sharing, discussion platforms, chat rooms, and polls. Video conferencing facilitates the connectivity of people, allowing teachers and students to gain access to online learning opportunities (Burke et al., 2022). Video conferencing assists in overcoming the communication and psychological distance between learners and teachers and allows them to interact and participate (Torrato et al., 2021). Therefore, teachers should utilise technology efficiently in online learning by leveraging all the engaging video features during video conferencing sessions to increase student engagement.

2.3. Teachers' movements and behaviour in video conferencing

Teachers' movements and behaviours in video conferencing significantly impact student engagement. According to Aelterman et al. (2019), teachers' motivating behaviours positively affect student engagement and feelings of competence. Teachers' behaviours provide autonomy support, students' sense of competence, and relatedness support (Aelterman et al., 2019; De Meester et al., 2020). These motivating behaviours include asking questions, encouraging problem-solving, providing constructive feedback, and developing mutually positive relationships with students (Aelterman et al., 2019; Haerens et al., 2013).

In addition to the aforementioned teacher behaviours that enhance engagement, previous research has indicated that certain body movements and non-verbal cues can also improve student engagement during video conferencing. One of the main obstacles to online learning that affects student engagement is the physical distance (Aladsani, 2021). Teachers use non-verbal cues such as eye gazing, silence, appropriate facial expressions, and appropriate body language to make video conferencing more effective (Jia et al., 2021; Trenholm et al., 2019).

An instrument is required to measure the engaging teaching videos according to teachers' movements and behaviours. However, most instruments currently available to measure student engagement are developed for face-to-face learning environments (Lee et al., 2019). Few researchers have investigated student engagement in online learning and developed assessment tools. Halverson and Graham (2019) stated that with clear indicators, student engagement could be measured and recommended a framework by providing indicators for cognitive and emotional engagement. Lee et al. (2019) also highlighted the importance of indicators and recognised them as an essential factor that positively affects learning and engagement. They emphasised the necessity of having an appropriate measurement tool for student engagement to ensure the quality of education. They developed an instrument with six factors with 24 items on a five-point Likert scale.

In the above section, the authors have established the importance of teachers' behaviours and movements in video conferencing and indicators to measure student engagement; however, no detailed list of characteristics and indicators of engaging teaching videos before the

Authors (2023) identified those in their previous study (phase 1). A systematic literature review was performed where authors identified 11 characteristics that can provide aid for enhancing student engagement via video conferencing. As shown in Table 1, the identified characteristics of engaging teaching videos have descriptive indicators and are categorised into three themes: Teachers' behaviours, movements, and use of technology). These identified indicators and characteristics are essential in improving student engagement.

Authors (2023) strongly believe that the characteristics and indicators stated in Table 1 above can be used as a benchmark to increase teachers' performance in online learning. Educational institutes can implement these indicators and characteristics of engaging teaching videos to enhance and regulate online teaching practices. This information can also be used by institutes around the world to create and provide training for teachers to enhance their skills in creating teaching videos in such a way that it enhances their students' engagement. In addition, Artificial Intelligence (AI) instruments can be used to identify the characteristics of student engagement in the online learning environment and later improve the teaching video qualities to incorporate engagement indicators.

2.4. AI and education

AI has been widely adopted in different applications, such as healthcare (Shaik, Tao, Higgins, et al., 2022; Tao et al., 2021) and education, with its ability to read and analyse text and videos (Shaik, Tao, Li, et al., 2022). In their study, Leisner et al. (2020) discussed three different learning conditions and explored the influence of in-video quizzes to annotate learning success and interest. The study was conducted on 78 students from four classes delivered by the same physics teacher. Wróblewska et al. (2022) proposed a framework to assess academic lectures based on quantitative features and provide objective feedback to help lecturers improve their didactic behaviours or course contents. In another study, the researchers (Liu et al., 2023) presented bidirectional encoder representations from transformers (BERT) model for cognitive presence identification. This AI model revealed the evolution and differences in MOOC learners' cognitive presence levels. The authors adopted AI methods such as computer vision and deep learning to process and annotate video lecture recordings. Gholamrezaie et al. (2022) proposed an intelligent system AI-EVL for effective learning by searching and enriching YouTube videos. The system visually informs the user about the contents of the video before watching it and then segments the video based on time slots and displays subtitles and annotated information.

Advancements in deep learning have fast-forwarded the process of image or video classification. Deep learning models include recurrent neural networks (RNN) (Jeong & Cho, 2022), convolutional neural networks (CNN) (Vrskova et al., 2020), and transformers (Liu et al., 2022). These models are a kind of neural network architecture that is effective in natural language processing tasks but has also been applied to computer vision tasks such as object detection and image recognition, which have the potential to learn the videos in the form of images and classify the videos. The technology field is abuzz with the latest development of ChatGPT, an AI-based tool created by OpenAI for generating text. This tool is designed to provide relevant responses and comprehend natural language to user queries. Within two months of its launch, it garnered a massive user base of over 100 million, prompting OpenAI to announce a subscription plan for unrestricted access and faster response times (Halaweh, 2023).

While ChatGPT can revolutionise educational activities, it poses certain concerns regarding potential bias (Navigli et al., 2023), privacy (Ali et al., 2023), job loss (Grassini, 2023), and academic integrity (Cotton et al., 2023; Perkins, 2023). The authors argue that teachers and learners must develop competencies to understand technology, limitations and unexpected vulnerabilities to incorporate large language models (Kasneci et al., 2023). ChatGPT is a versatile educational tool,

Table 1
Main theme, characteristics, and indicators of engaging teaching videos (Authors, 2023, p. 11, p.11).

Main theme	Characteristics	Indicators
Teachers' Behaviours	Encourage Active Participation	• Encouraging students' participation in discussion • Encouraging students to share their knowledge and ideas • Encouraging students to ask questions • Encouraging collaborative learning activities • Encouraging meaningful interaction • Encouraging students to turn on their webcams
	Establishing Teacher Presence	• Clear and concise explanations of information • Recognising and considering learners' Individual differences • Using an appropriate style of presentation • Allowing sufficient time for students' information processing • Providing Learning resources • Giving clear instructions • Using a range of teaching strategies • Appropriate speed of lecture delivery
	Establishing Social Presence	• Maintaining constant teacher-student interaction • Encouraging student-student interaction (Peer collaboration) • Active and constructive communication • Taking on multiple roles
	Establishing Cognitive Presence	• giving students a sense of puzzlement (trigger) • providing opportunities for students to reflect (exploration) • leading students to think and learn through discussion with others (integration) • helping students apply knowledge to solve issues (resolution)
	Questions and Feedback	• Addressing students' questions & Providing prompt feedback • Asking for questions and feedback • Clarifying misunderstanding
	Displaying Enthusiasm	• Motivating students • Displaying positive emotion
	Establishing Clear Expectations	• Outlining the learning objectives • Outlining teachers' expectations of students' behaviours and responsibilities
	Demonstrating empathy	• Using appropriate changes in tone of voice • Ensuring the learning environment is a respectful, safe, and supportive one • Showing concern
	Demonstrating Professionalism	• Demonstrating in-depth and up-to-date knowledge • Displaying appropriate behaviours
Teachers' Movements	Using non-verbal cues	• facial expressions • gestures • eye gazes • silence • eye contact • physical proximity • appropriate body language
Use of Technology	Using technology effectively	• Screen sharing & Enabling Chat, Camera, and Microphone • Varying the presentation media • Providing technical support to students • Providing multiple communication channels • Providing interactive software tools • Enabling class recording for later review

offering teachers intelligent assistance in grading and language support while providing learners with an interactive and adaptive learning experience that fosters creativity and engagement. Careful implementation is required to navigate ethical concerns and potential biases, and the integration of ChatGPT should complement traditional teaching methods to ensure a comprehensive educational experience. A pedagogical approach that prioritises critical thinking and fact-checking and a well-defined strategy within educational systems are essential to incorporate and fully leverage the capabilities of extensive language models in teaching and learning settings.

In this context, Zhai (2022) suggests adjusting learning goals to include using AI tools for subject-domain tasks, emphasising creativity and critical thinking, and designing AI-based learning tasks to assist students in solving problems. Furthermore, there is a need for new assessment formats that focus on skills that AI cannot replace. The latest version of ChatGPT, GPT-4, has multimodal learning and generation capabilities, including the ability to analyse videos.

Mubarak, Cao, and Ahmed (2020) conducted sequential temporal classification by analysing video clickstream data, which can predict the performance of learners and address their issues to improve the educational process. The authors deployed a variant of the RNN model known as long short-term memory (LSTM) on characteristics derived from video data to forecast weekly learner performance and assist teachers in setting measures for timely intervention. The LSTM model outperforms the other baseline models' logistic regression (LR), artificial neural networks (ANN), and support vector machine (SVM) with an accuracy rate of 93%. Hieu et al. (2021) proposed an automated system that allows schools to capture entire sessions and summarise students' behaviour in the classroom. The authors used the deep neural networks (DNN) model and trained it with 1.2 million images, achieving an accuracy of 88.9%. Mubarak, Cao, Zhang, and Zhang (2020) proposed an LSTM model-based visualisation tool to address the research question, "How does learners' behaviour in videos impact their performance in the MOOC course?". The deep learning was trained with Massive Open Online Courses (MOOCs) course videos as a time series sequence. The model achieved an accuracy of 90% in predicting the learners' performance and enabling teachers to take timely actions for intervention.

Bhatti et al. (2021) provided a feedforward learning model that can assess the facial expressions of an instructor in a classroom. The authors extracted features using the CNN model and employed the Regularized Extreme Learning Machine (RELM) model to classify five different expressions as amusement, awe, confidence, disappointment, and neutral of the instructor within the classroom. The proposed model can achieve the best performance of 96.8% compared to other baseline models.

In conclusion, numerous studies have concluded that teachers should utilise appropriate behaviours and movements in online learning to increase their presence, which enhances student engagement (Cents-Boonstra et al., 2021; Dewan et al., 2019). The identified teachers' behaviours and movements could aid in training deep learning algorithms. Nonetheless, a notable research gap exists in the field of AI, as there is currently no AI instrument capable of identifying the indicators of engaging teaching videos. The development of an AI tool proficient in discerning these specific engaging indicators is an unexplored opportunity. This tool has the potential to greatly aid educators and educational institutions in enhancing the effectiveness of learning and enriching the overall educational journey for students.

3. Research gaps

Drawing from the analysis of existing literature, a conspicuous research gap emerges. Presently, no established video annotation procedure exists to assist AI engineers in training AI-enabled tools that can support in improving the teaching and learning process. Moreover, a notable absence persists in terms of an AI tool capable of generating comprehensive reports on the indicators and characteristics of engaging teaching videos.

4. Methodology

The authors have employed a designed-based research (DBR) approach to develop an AI tool that generates a report whenever a video recording is analysed for teachers' behaviours and movements.

Researchers (Barab & Squire, 2004; Oh & Reeves, 2010; Van et al., 2006) used various terms such as design experiments, development research, educational design research, and design research; however, design-based research became the dominant one used for the research paradigm that is used for creating educational technologies (Miah et al., 2020). The usage of DBR increased significantly, mainly with technological innovation and interventions in education (Anderson & Shattuck, 2012). This approach allows researchers to produce tools, approaches, and theories (McKenney & Reeves, 2018).

The authors structured this study in three phases to design an Artificial Intelligence tool. The first phase (Authors, 2023) identified the teachers' behaviours and movements in video conferencing as characteristics and indicators of engaging teaching videos. In this study, the authors identified 47 indicators and 11 characteristics categorised into three main themes (see Table 1).

This current study focuses on the second phase of DBR (prototyping), which involves video annotation to create an AI-enabled tool. Authors have applied the identified indicators and characteristics of engaging teaching videos to recorded lecture videos using Zoom to design an AI-enabled that autogenerate a report on engaging teaching videos. The tool is designed through two prototypes. This study also explains how these indicators and characteristics were applied in training deep learning algorithms, which is a classification tool for annotating videos. In this stage, an AI expert has assisted in creating an AI-enabled instrument. In the last phase of the DBR research, which will be reported in another study, the authors will compare the performance of the AI methodology annotation to manual annotation and evaluate the entire process to further enhance the deep learning instrument.

4.1. Data collection and analysis

The process of designing the AI-enabled tool is illustrated in Fig. 1, which outlines the process starting from the video collection process and manual annotation using the VIA tool. With the assistance of an AI expert, the annotated videos are pre-processed through several tasks, including splitting the videos into chunks, converting chunks into images, labelling images, and dividing them into training and evaluating data for deep learning model training and evaluation. The next step involves implementing prototype 1 for modeling, evaluating, and classification results from an educational and AI perspective. The challenges identified in data and modeling led to the development of prototype 2, which oversamples the data to overcome these challenges. This section presents two sets of results, one for each prototype in the result section.

4.2. Video collection process

The authors gathered the recorded lecture videos, which are recordings of lectures presented by a regional university of Australia's teachers to higher education students. The lecturers have used Zoom as the video conferencing software while presenting their lectures. Based on the discussion with the experts in AI, a data size of 25 recorded lecture videos has been selected. These recorded lecture videos have been presented to higher education students and cover health, law, business, education, engineering, sciences, and arts disciplines. The video duration ranges from 00:59:06 to 01:51:52 with an average time of 01:28:37. In the selection process, the authors ensured to select various presenter settings such as presenter location on the top right corner, middle, and bottom right corner. We included poor, average, and good-quality videos to train the deep learning model in various settings. The authors also paid attention to other variabilities while selecting the videos, such as videos where students also turned on their cameras and

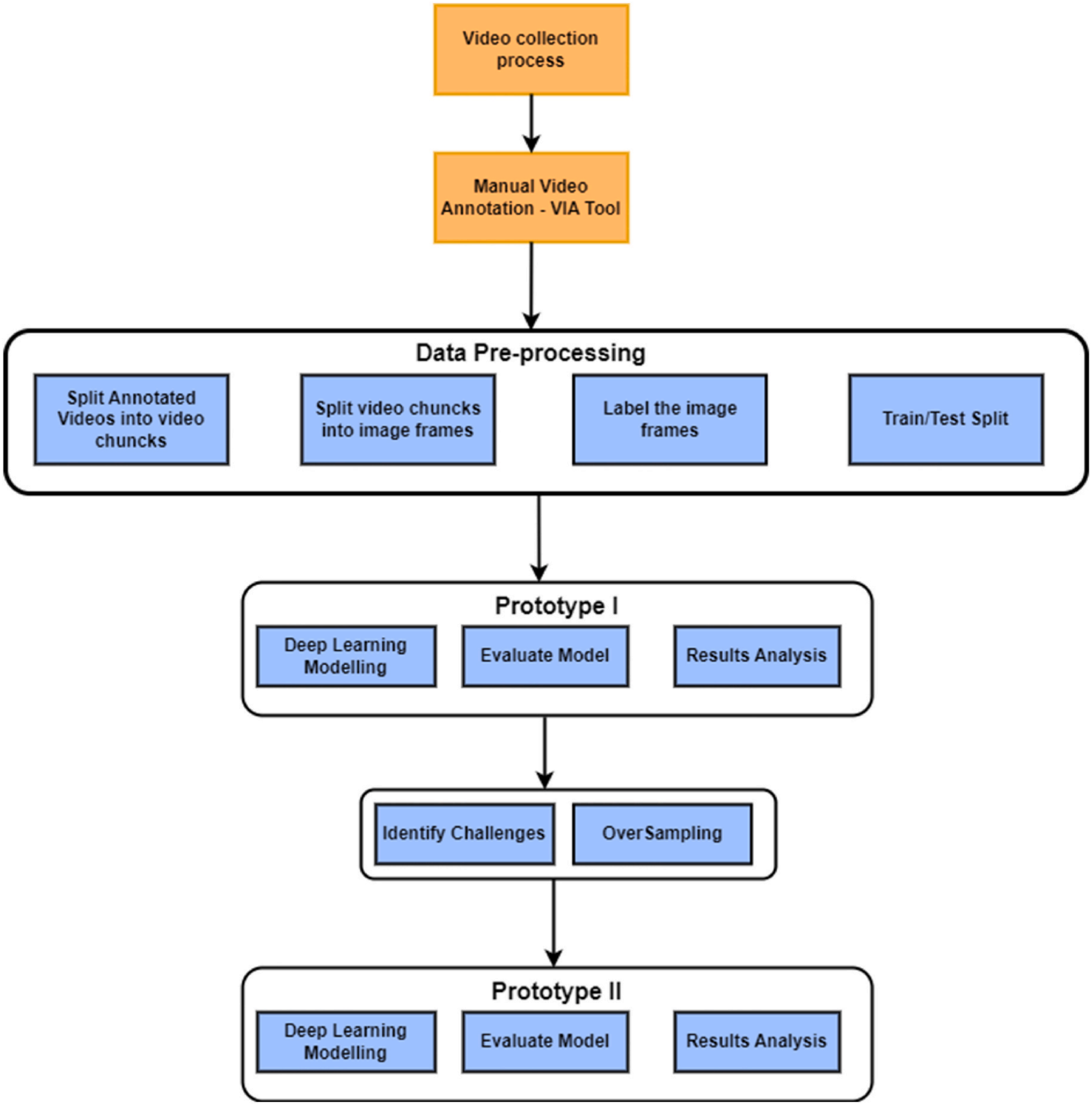

Fig. 1. Experimental design of the proposed methodology.

videos where the presenter's camera location changed while presenting. The authors gathered and stored the data (lecture videos) in cloud storage (OneDrive) and categorised it into demographics, with 13 female and 12 male presenters. The authors have ethics approval from the University of Southern Queensland with ethics approval number H20REA185. The authors have not gathered information regarding the lecturers' age, location and academic background.

4.3. Manual annotation of videos

The authors manually annotated 25 recorded lecture videos individually. In this manual annotation process, the authors have used VGG Image Annotator (VIA) software to annotate zoom-based lecture recordings. VIA is an open-source project-based annotation software for annotating images, audio, and videos available at https://www.robots.ox.ac.uk/~vgg/software/via/app/via_video_annotator.html. The manual annotation has been performed for each video on the indicator level. All indicators, characteristics and their main themes are shown in Table 1, and deep learning has been trained on the indicators level. The exported reports from this manual annotation assisted the AI engineer in training the AI tool. The processes of manual annotation is attached separately to this paper.

4.4. Involvement of AI expertise

The authors enlisted the support of an AI expert to develop an AI tool capable of producing a report based on certain indicators and

Table 2
AI expert involvement.

Process	Involvement
AI process	Extracting temporal coordinates from videos and transforming them into image frames to train convolutional neural networks (CNN) model
Data pre-processing	Captured the annotated image frames for the convolution layer of the deep learning model
Deep learning model	Developed the CNN model as a deep learning approach
Model evaluation	Collaborating with the other authors in performing the Model evaluation

characteristics of engaging teaching videos. The AI expert played a role in the processes mentioned in Table 2, further described in the subsequent section.

4.5. AI process

In AI methodology, the authors focused on creating a deep learning model to learn a teacher's actions in a recording with the support of an AI expert. This is achieved by recording the temporal coordinates extracted from the tool's manual video annotation. Temporal coordinates are markers in the video timeline that help identify specific points in time. Selected lecture videos split based on these coordinates and transformed them into a stack of image frames, as depicted in Fig. 2. The pre-processed frames were then labelled with corresponding teaching indicators and prepared the data model for training. Next, the data was split into two sets - train and test - for model training and evaluation. The training set was used to make the deep learning model learn the frames and corresponding labels. The testing set was used to evaluate how well the model was learned. An AI expert fed the train set to the CNN model to learn the actions in image frames and their corresponding label. Finally, the test set was used to assess the effectiveness of the CNN model.

4.6. Data pre-processing

During the data pre-processing step, the AI expert captured the temporal coordinates provided by the video annotation tool. For example, suppose a lecture recording displays the teaching indicator "Clear and concise explanation of information" at the temporal coordinates (3051.315, 3053.256). In that case, the recorded lecture was divided into video segments highlighting and extracting the teaching indicator. Then, each video was split into segments into image frames and annotated each frame with the "Clear and concise explanation of information" teaching indicator. These annotated image frames are represented as 3D matrices and serve as input for the convolution layer of the deep learning model, as described in the subsequent subsection.

4.7. Deep learning model

The AI expert utilised and developed the CNN model as deep learning to classify images that contain three-dimensional (3-D) data, encompassing width, height, and colour channels (such as RGB). CNN model was selected for this study due to its ability to automatically and adaptively learn spatial hierarchies of features from the input data, making them particularly suitable for our study involving video classification, where spatial features play a vital role. CNNs have consistently demonstrated state-of-the-art performance in various image and video recognition tasks, providing strong motivation for their application in our study (Hasnine et al., 2021; Pabba & Kumar, 2021; Sharma et al., 2022). The flexible architecture of CNNs allowed us to design and optimise the network to suit best the specific characteristics and challenges of the teaching behaviours we aimed to classify. Additionally, CNNs are scalable to larger datasets, allowing for future expansion of the research should more data become available. The strength of the CNN model lies in its ability to manage the high dimensionality of images by preserving essential information while compacting the overall data structure. Fig. 3 portrays the learning workflow of the CNN model. Initially, the pre-processed input image frames, containing both spatial dimensions and colour information, are sent to a two-dimensional (2D) convolution layer. This layer applies various filters to break down the image into smaller sub-images, allowing the model to examine specific features within these sections. Following the convolution process, the pooling layer receives the output, selecting the most significant value from each feature group and crafting a down-sampled representation of the features. This down-sampling technique aids in reducing computational complexity and enhancing model robustness. The pooled features are then flattened into a 2D array to be further processed in the CNN's output layer. This final layer computes a probability for each potential classification label, and a specific threshold can be set to translate these probabilities into definitive class labels.

4.8. Model evaluation

The AI expert collaborates with the other authors in the model evaluation. The evaluation process considered six metrics, and each is described below.

Accuracy: This measures the proportion of labels accurately generated by the model compared to all the labels.

Precision: This measures the proportion of labels precisely generated by the model among all labels generated by the model.

Recall: This measures the proportion of labels accurately generated by the model among all the labels present in the video.

F1-score: This is a combined measure of precision and recall and is commonly used to assess the overall performance of the model.

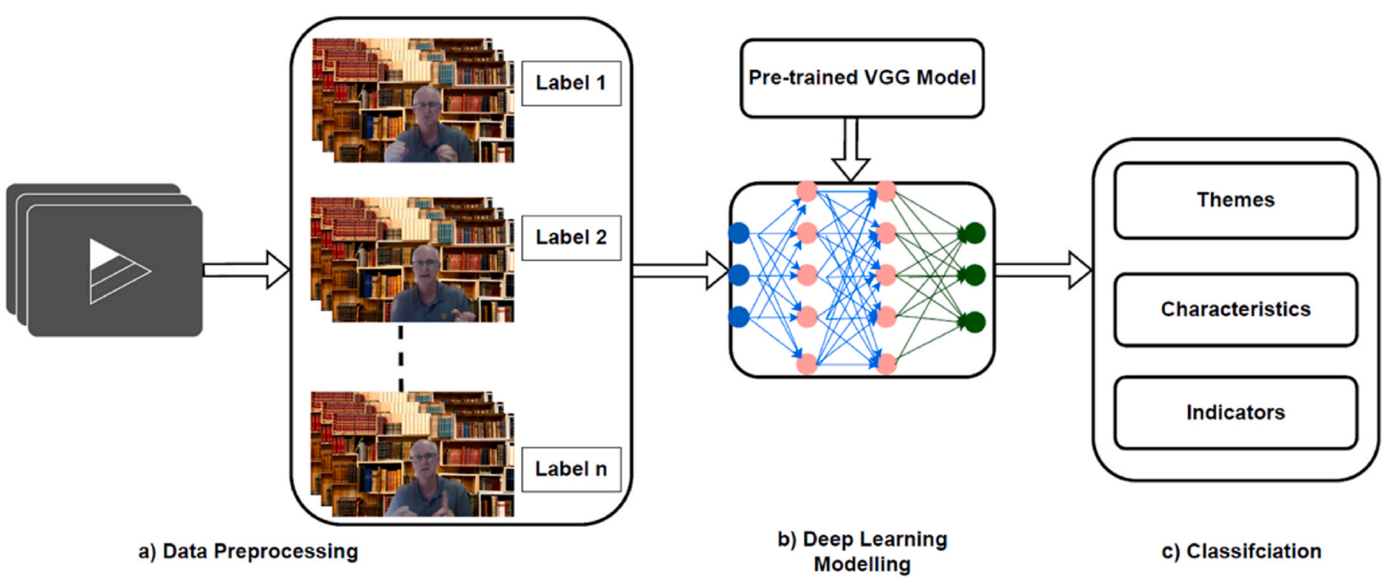

Fig. 2. Proposed AI methodology for video annotation.

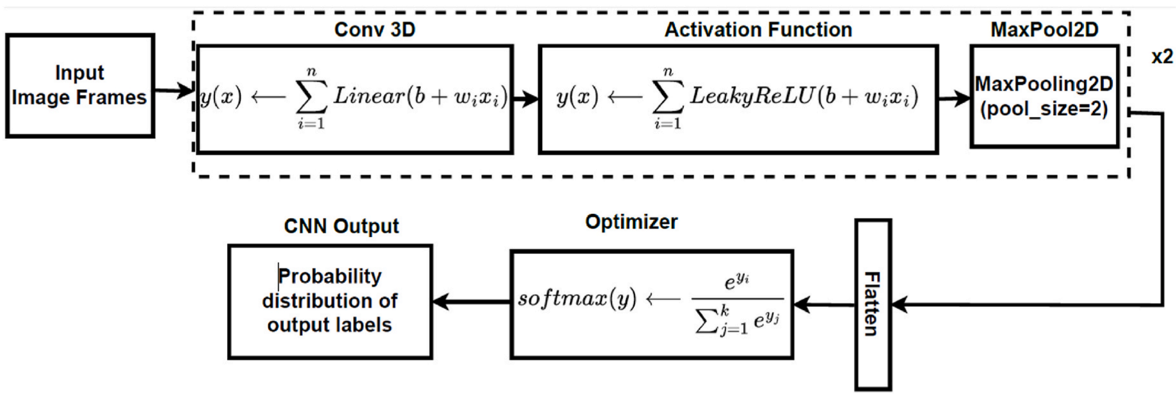

Fig. 3. Proposed CNN model learning process.

Cohen's Kappa: Cohen's Kappa measures the agreement between two raters, adjusting for the probability of random agreement. It ranges from −1 (complete disagreement) to 1 (complete agreement), with 0 indicating agreement by chance.

Area Under the Curve (AUC): AUC represents the area under the curve, providing a threshold-independent evaluation of a binary classifier's ability to distinguish between classes. It ranges from 0 to 1, with 0.5 representing no discrimination and 1 representing perfect discrimination.

5. Results

In this section, the authors present the outcomes of prototype 1, where a model is built and prototype 2, where the model is refined further. Firstly, the authors discuss the challenges identified in prototype 1, which facilitated its improvement and the subsequent development of prototype 2. Additionally, the authors thoroughly examined prototype 2 as a potential key to these challenges and discussed the results obtained.

5.1. Model building: prototype 1

In this study, the video annotation is conducted at different levels of labels on the pre-processed videos. The labels are categorised into themes, characteristics, and indicators. In this case, the themes are Teachers' behaviours, Teachers' movements, and the use of technology. Six evaluation methods are used to report the results where precision measures the true positive predictions (i.e., when the model correctly identifies a video as belonging to a certain theme) among all positive predictions. If the precision is high, it signals that the model is making many false positive predictions. Recall measures the true positive predictions, and a high recall means that the instrument correctly identifies most of the videos that belong to a certain theme. The F1-score is a metric of a model's accuracy that considers both precision and recall and is a commonly used metric for classification tasks. Balanced accuracy is the average recall obtained in each class. This metric looks at the recall for each class individually and gives an overall accuracy score. Cohen's Kappa is used to measure the agreement between two raters, adjusting for the probability of random agreement. Area Under the Curve (AUC) represents the area under the curve, providing a threshold-independent evaluation of a binary classifier's ability to distinguish between classes.

Table 3 comprehensively evaluates three themes: Teachers' Behaviours, Teachers' Movements, and Use of Technology in the context of video classification. For Teachers' Behaviours, the model shows a balanced performance with precision, recall, and F1-score, all at 0.74, indicating a consistent ability to identify this theme accurately. The Balanced Accuracy of 0.75 and Cohen's Kappa score of 0.73 reinforce this balanced performance, while the AUC of 0.85 demonstrates excellent discriminatory power. The theme Teachers' Movements shows slightly improved precision at 0.78 but maintains a similar recall at 0.74, resulting in an F1-score of 0.76. The Balanced Accuracy and Cohen's Kappa values are 0.79 and 0.75, respectively, indicating good overall classification and substantial agreement. The AUC value of 0.81 further illustrates a strong capability to differentiate classes. Use of Technology, on the other hand, exhibits a lower precision of 0.64 but compensates with a higher recall of 0.76, reflecting a tendency to accurately capture positive instances at the potential expense of false positives. The F1-score of 0.70, the exceptional Balanced Accuracy of 0.95, and the AUC of 0.93 indicate the model's superior ability to classify and discriminate this theme despite its lower precision. Cohen's Kappa score of 0.68 denotes substantial inter-rater agreement.

Table 4 presents the classification performance across various

Table 4
Characteristic-Level performance of the CNN model.

Characteristics	Precision	F1-Score	Recall	Balanced Accuracy	Cohen's Kappa	AUC
Encourage Active Participation	0.77	0.76	0.77	0.85	0.74	0.83
Establishing Teacher Presence	0.67	0.75	0.70	0.82	0.66	0.79
Establishing Clear Expectations	0.60	0.67	0.63	0.79	0.58	0.74
Demonstrating Empathy	0.61	0.28	0.39	0.76	0.45	0.71
Using Non-verbal Cues	0.75	0.86	0.80	0.81	0.73	0.82
Using Technology Effectively	0.78	0.61	0.68	0.87	0.71	0.88

Table 3
Theme-Level performance of the CNN model.

Themes	Precision	Recall	F1-Score	Balanced Accuracy	Cohen's Kappa	AUC
Teachers' Behaviours	0.74	0.74	0.74	0.75	0.73	0.85
Teachers' Movements	0.78	0.74	0.76	0.79	0.75	0.81
Use of Technology	0.64	0.76	0.70	0.95	0.68	0.93

educational characteristics. In encouraging active participation, the model performs well with precision, recall, and F1-score, all around 0.76–0.77 and an AUC of 0.83. Establishing teacher presence has fair outcomes with precision at 0.67, an F1-score of 0.75, and an AUC of 0.79, although the small data distribution may affect reliability. The model struggles more with establishing clear expectations and demonstrating empathy, with inconsistencies in precision, recall, and F1-scores and lower Cohen's Kappa values, reflecting moderate agreement. However, it excels in using non-verbal cues and technology effectively, especially in the high AUC values of 0.82 and 0.88, respectively. Overall, the model exhibits varying success across different characteristics, performing strongly in some areas while facing challenges in others, and the different data distributions might also impact the generalizability of these results.

Table 5 shows performance results for various teaching indicators in a multi-label classification task. The indicators are listed in the first column, and the corresponding results for precision, recall, f1-score, balanced accuracy, Cohen's Kappa, AUC and data distribution are shown in the following columns. The data distribution column shows the number of instances each indicator is present in the data used to analyse the performance of the multi-label classifier. The precision of an indicator is correctly predicted positive instances divided by the sum of true positive instances and incorrectly predicted positive instances (false negative instances). The recall of an indicator is the number of true positive instances divided by the sum of true positive instances and false negative instances (incorrectly predicted negative instances). The f1-score is a weighted average of precision and recall and is a frequently used metric for evaluating the performance of multi-label classifiers. The balanced accuracy is the average for each class, considering both the positive and negative classes. Cohen's Kappa measures the agreement between two raters, adjusting for the probability of random agreement. It ranges from −1 (complete disagreement) to 1 (complete agreement), with 0 indicating agreement by chance. AUC represents the area under the ROC curve, providing a threshold-independent evaluation of a binary classifier's ability to distinguish between classes. It ranges from 0 to 1, with 0.5 representing no discrimination and 1 representing perfect discrimination.

The zero metrics indicate a critical issue rooted in data distribution and have significant implications for AI models. This phenomenon arises when there is a severe class imbalance, where certain classes have a disproportionately smaller number of instances compared to others within the training dataset. In our case, the data distribution column reflects this smaller number of instances issue. This class imbalance issue is widely recognised in the field of machine learning and has been extensively studied due to its far-reaching consequences (Johnson & Khoshgoftaar, 2019).

From the results, it can be seen that the teaching indicators with higher f1-scores, precision, recall, AUC, and Cohen's Kappa values are: "Encouraging students to share their knowledge and ideas" (f1-score: 0.75, precision: 0.7, recall: 0.79, AUC: 0.8, Cohen's Kappa: 0.7), "Outlining the learning objectives" (f1-score: 0.63, precision: 0.61, recall: 0.65, AUC: 0.75, Cohen's Kappa: 0.6), and "appropriate body language" (f1-score: 0.73, precision: 0.72, recall: 0.75, AUC: 0.81, Cohen's Kappa: 0.74). The model has effectively predicted these indicators. On the other hand, indicators with zero scores in precision, recall, f1-score, Cohen's Kappa, and AUC of 0.5, such as "Encouraging students' participation in discussion," "Enabling class recording for later review," "Providing Learning resources," "Encouraging students to ask questions," "Encouraging collaborative learning activities," "eye contact," "Varying the presentation media," and "Encouraging meaningful interaction," have not been well-predicted by the multi-label classifier. These categories failed due to limited data availability or possibly challenges inherent in classifying these specific behaviours.

5.2. Identifying challenges

The analysis of prototype 1 revealed two major issues: an imbalanced dataset and misleading metrics. Addressing these issues is essential for developing an accurate tool for classifying teaching videos. To overcome these challenges, the AI expert developed a prototype, prototype 2, which implemented solutions to both issues.

Table 6 provides a detailed explanation of each problem and outlines the solutions employed by the AI expert to resolve them. The challenges faced in training the prototype 1 are listed in the first column. A detailed explanation of the challenges is stated in the second column, and the solution adopted by the AI expert in the last column.

Table 6
Challenges identified in prototype 1 and the solution adopted.

Challenge	Explanation	Solution Adopted
Imbalanced dataset	An imbalanced dataset is one where the number of instances in the target classes is unequal, leading to an unequal distribution of the target variable. This can lead to a biased model that is successful for the majority but less for the minority.	Oversampling addresses this problem by duplicating samples from the underrepresented class in the dataset until the class distribution is balanced.
Misleading Metrics	Oversampling aims to address this problem by duplicating samples from the underrepresented class in the dataset until the class distribution is balanced.	Alternative evaluation metrics are adopted to overcome this issue.

Table 5
Indicator-level performance of the CNN model.

Indicators	precision	recall	f1-score	Balanced Accuracy	Cohen's Kappa	AUC	Data Distribution
Encouraging students' participation in discussion	0.00	0.00	0.00	0.50	0.00	0.50	7
Enabling class recording for later review	0.00	0.00	0.00	0.50	0.00	0.50	5
Providing Learning resources	0.00	0.00	0.00	0.50	0.00	0.50	2
Giving clear instructions	0.41	0.69	0.52	0.70	0.40	0.65	37
Encouraging students to share their knowledge and ideas	0.70	0.79	0.75	0.82	0.70	0.80	152
Encouraging students to ask questions	0.00	0.00	0.00	0.49	0.00	0.50	19
Outlining the learning objectives	0.61	0.65	0.63	0.80	0.60	0.75	36
Using appropriate changes in tone of voice	0.55	0.50	0.52	0.75	0.53	0.72	94
Facial expressions	0.52	0.24	0.33	0.73	0.35	0.64	57
Encouraging collaborative learning activities	0.00	0.00	0.00	0.49	0.00	0.50	10
Eye contact	0.00	0.00	0.00	0.50	0.00	0.50	1
Appropriate body language	0.72	0.75	0.73	0.78	0.74	0.81	293
Encouraging meaningful interaction	0.00	0.00	0.00	0.50	0.00	0.50	4
Screensharing & EnablingChat, Camera, and Microphone	0.41	0.52	0.46	0.69	0.45	0.68	86
Varying the presentation media	0.00	0.00	0.00	0.50	0.00	0.50	5

5.3. Model refinement: prototype 2

The outcomes of the second prototype are described in this section, developed after identifying the challenges encountered in prototype 1 and taking appropriate measures to overcome them in prototype 2. Prototype 2 is the refined version of prototype 1. The pre-processed data was fed into the deep learning model, and training and evaluation were conducted. The outcomes of prototype 2 are presented and analysed in this section.

5.4. Theme level results

The authors evaluate and compare the performance of the artificial intelligence model on the three different themes before and after the improvements in data.

As shown in Table 7, the evaluation of three distinct themes in teaching—Teachers' Behaviours, Teachers' Movements, and Use of Technology—reveals significant insights into the model's performance. For the theme of Teachers' Behaviours, the model exhibits a precision of 0.81, recall of 0.82, F1-score of 0.84, balanced accuracy of 0.85, Cohen's Kappa of 0.82, and AUC of 0.87. In the case of Teachers' Movements, the metrics are even higher, with precision, recall, F1-score, balanced accuracy, Cohen's Kappa, and AUC values of 0.82, 0.83, 0.85, 0.89, 0.84, and 0.89, respectively. Finally, for the theme of Use of Technology, the values are 0.75 for precision, 0.86 for recall, 0.81 for F1-score, 0.88 for balanced accuracy, 0.78 for Cohen's Kappa, and 0.86 for AUC. Comparing the two tables indicates that the improved results have higher values for all evaluation metrics across all three themes. This improvement suggests that the refined model or hyperparameters have enhanced the dataset's accuracy and ability to classify these themes. However, it is essential to note that the data distribution for each theme is uniform at 361, which might impact the interpretation of these results. In scenarios where data distribution varies significantly, it may affect the model's performance, especially if the dataset is imbalanced.

5.5. Characteristic level results

Table 8 presents the characteristic-level performance of the improvised CNN model, detailing the precision, recall, F1-score, Cohen's Kappa, AUC, balanced accuracy, and data distribution for six different characteristics related to teaching.

The performance for "Encourage Active Participation" is notable with a precision of 0.81, recall of 0.87, F1-score of 0.82, Cohen's Kappa of 0.82, AUC of 0.86, and balanced accuracy of 0.86. Similarly, "Establishing Teacher Presence" shows strong results, particularly in precision (0.82) and AUC (0.84). "Establishing Clear Expectations" demonstrates a promising F1-score of 0.76, while "Demonstrating Empathy" has an impressive precision of 0.84. The characteristics "Using Non-Verbal Cues" and "Using Technology Effectively" also display robust outcomes, with the latter achieving a remarkable AUC of 0.85.

A significant observation from this table is the consistent data distribution across all characteristics (351 for each). This uniform distribution indicates that the dataset has been balanced, likely through oversampling or other data-level enhancements. The balanced distribution and improvements across multiple evaluation metrics signal that the model performs better in classifying these teaching characteristics. This improvement enhances the model's accuracy and reliability. It suggests that the CNN model is now better poised to make predictions on new or unseen data, reflecting a more faithful representation of the underlying patterns in the teaching domain.

5.6. Indicator level results

Similarly, balancing the data appears to have refined the overall performance of the indicators (see Table 9), as indicated by the increased precision, recall, F1 score, Cohen's Kappa, and AUC across almost all indicators (see Table 10).

In summary, prototype 2 achieved better performance than prototype 1 by addressing the identified challenges. It demonstrated improved efficiency in classifying videos based on three levels of labelling: themes, characteristics, and indicators.

6. Baseline model comparisons

In this section, we present a comprehensive comparison between the proposed CNN architecture and traditional baseline models, including the k-Nearest Neighbors (k-NN) Classifier (Bourguet et al., 2020), Decision Tree (Zaletelj & Košir, 2017), and Support Vector Machine (SVM) (Thomas & Jayagopi, 2017). The comparison aims to demonstrate the efficacy of the proposed CNN model against conventional algorithms in classifying teaching behaviours.

The results of the classification models applied to identify teacher behaviours in video classification present a clear comparison of their effectiveness. Among the baseline models, the k-NN classifier shows a balanced performance with a precision of 0.72, recall of 0.7, and F1-score of 0.71. Its Cohen's Kappa score of 0.42 and AUC of 0.71 further highlight a reasonable level of agreement and discriminatory power. The Decision Tree model performs slightly worse, with all metrics around the 0.68 mark, reflecting a more modest performance in classification. Cohen's Kappa score of 0.36 also indicates a lower level of agreement between raters. The SVM model shows an improvement, especially in precision (0.75) and AUC (0.75), showing a better balance between classifying the positive and negative classes. However, our proposed CNN model significantly outperforms all the baseline models in all the metrics. With a precision of 0.82, recall of 0.8, and F1-score of 0.81, it demonstrates a superior balance between sensitivity and specificity in classifying teacher behaviours. Cohen's Kappa score of 0.62 signifies a substantial agreement, and the AUC of 0.82 indicates the excellent ability to distinguish between different teacher behaviours. The CNN model's robust performance explains its effectiveness in recognising and categorising teacher behaviours in video data. It is a promising tool for enhancing the analysis and understanding of pedagogical practices.

7. Discussion

This study has developed an AI-enabled tool to identify the teachers' behaviours and movements in engaging teaching videos. Teachers should use engagement-enhancing behaviours and movements in video conferencing to improve student engagement in online education settings. The authors have identified these teachers' behaviours and movements in phase 1 of the DBR project and established their importance in enhancing students' engagement. However, there is a need for an AI-enabled instrument to identify the indicators and characteristics of engaging teaching videos and generate a report. Thus, an instrument is

Table 7
Theme-Level performance of the improvised CNN model.

Themes	Precision	Recall	F1-Score	Balanced Accuracy	Cohen's Kappa	AUC	Data Distribution
Teachers' Behaviours	0.81	0.82	0.84	0.85	0.82	0.87	361
Teachers' Movements	0.82	0.83	0.85	0.89	0.84	0.89	361
Use of Technology	0.75	0.86	0.81	0.88	0.78	0.86	361

Table 8
Characteristic-Level performance of the improvised CNN model.

Characteristics	Precision	Recall	F1-Score	Cohen's Kappa	AUC	Balanced Accuracy	Data Distribution
Encourage Active Participation	0.81	0.87	0.82	0.82	0.86	0.86	351
Establishing Teacher Presence	0.82	0.75	0.79	0.81	0.84	0.84	351
Establishing Clear Expectations	0.71	0.7	0.76	0.75	0.81	0.82	351
Demonstrating Empathy	0.84	0.65	0.62	0.78	0.80	0.8	351
Using Non-Verbal Cues	0.79	0.87	0.88	0.84	0.86	0.83	351
Using Technology Effectively	0.82	0.78	0.72	0.79	0.85	0.86	351

Table 9
Indicator-Level performance of the improvised CNN model.

Indicators	Precision	Recall	F1-Score	Balanced Accuracy	Cohen's Kappa	AUC	Data Distribution
Encouraging students' participation in discussion	0.68	0.70	0.71	0.86	0.71	0.85	293
Enabling class recording for later review	0.70	0.75	0.69	0.81	0.72	0.80	293
Providing Learning resources	0.65	0.75	0.75	0.68	0.70	0.75	293
Giving clear instructions	0.61	0.75	0.60	0.75	0.67	0.76	293
Encouraging students to share their knowledge and ideas	0.75	0.82	0.85	0.86	0.79	0.87	293
Encouraging students to ask questions	0.65	0.69	0.71	0.85	0.70	0.83	293
Outlining the learning objectives	0.62	0.85	0.83	0.82	0.73	0.84	293
Using appropriate changes in tone of voice	0.65	0.70	0.60	0.75	0.67	0.77	293
facial expressions	0.62	0.72	0.80	0.79	0.70	0.80	293
Encouraging collaborative learning activities	0.68	0.69	0.71	0.78	0.70	0.79	293
eye contact	0.75	0.78	0.68	0.72	0.74	0.76	293
appropriate body language	0.78	0.85	0.83	0.80	0.79	0.82	293
Encouraging meaningful interaction	0.65	0.75	0.65	0.80	0.70	0.80	293
Screensharing & Enabling Chat, Camera, and Microphone	0.61	0.75	0.66	0.82	0.68	0.81	293
Varying the presentation media	0.78	0.69	0.81	0.82	0.76	0.83	293

The balanced accuracy has also improved significantly, indicating better overall model performance distinguishing between the two classes.

Table 10
Baseline models comparison.

Model	Precision	Recall	F1-Score	Balanced Accuracy	Cohen's Kappa	AUC
k-NN	0.72	0.70	0.71	0.71	0.42	0.71
Decision Tree	0.68	0.67	0.67	0.68	0.36	0.68
SVM	0.75	0.73	0.74	0.74	0.48	0.75
CNN (Ours)	0.82	0.80	0.81	0.81	0.62	0.82

required to reflect the teachers' behaviours and movements in engaging teaching videos.

7.1. Exploration of research findings

In this research, the authors identified that most instruments currently available to monitor student engagement are designed for in-person learning environments. Only a few instruments are available that can predict student engagement in online learning settings; however, they do not provide clear indicators for engaging teaching videos. The authors strongly believe that indicators are required to observe and measure engaging teaching videos. Researchers (Halverson & Graham, 2019; Lee et al., 2019) have developed instruments with indicators to predict student engagement in online learning environments; however, the indicators measuring student engagement are broad and do not measure it based on teachers' behaviours and movements. Therefore, the authors have performed a systematic literature review in phase 1 of this DBR research project and identified 11 characteristics and 47 descriptive indicators (Authors, 2023). These behaviours and movements are essential to enhance student engagement (Aelterman et al., 2019; Aladsani, 2021; De Meester et al., 2020; Jia et al., 2021). Prior use of instruments to measure student engagement required ongoing manual human analysis, which is inherently biased, so this paper has presented a method that uses artificial intelligence to reduce such bias. With the assistance of an AI expert, the authors created an AI-enabled instrument that can automatically identify the indicators and characteristics of engaging teaching videos and generate a report.

In the current study, the authors employed the DBR methodology to design the AI-enabled tool, as this approach is suitable for technological interventions. Video analysis with Artificial Intelligence was employed to address the gap, specifically using a deep learning model called convolutional neural networks (CNN). The model was applied to video features to classify teaching characteristics, themes, and indicators. This approach aims to assist teachers in identifying areas for timely intervention.

Following the guidelines for DBR, the tool is designed through two prototypes. An AI expert assisted in creating the prototypes, where the authors observed two issues in the results of prototype 1. The first issue was an imbalanced dataset where the number of instances in the target classes was unequal, leading to an unequal distribution of the target variable. This can result in a bias, where the model might be able to perform successfully for the majority class but unsatisfactory for the minority class. Based on the results discussed in the prototype I, the classification model can perform at themes and characteristic levels of the data due to the availability of more data. However, at the indicator level, the model performance is poor. The data distribution column in Tables 4 and 6 show the imbalance in the number of records among the labels. An oversampling technique is adopted in data modeling to overcome the data imbalance issue. Oversampling is a technique used in artificial intelligence to balance class distribution in the dataset (Shaik, Tao, Li, et al., 2022). In a multi-class classification problem, class imbalance occurs when one class has significantly more samples than the others. This can cause the classifier to perform poorly on the underrepresented class as it may be biased towards the majority class. Oversampling addresses this problem by duplicating samples from the underrepresented class in the dataset until the class distribution is balanced. This process improved the performance of the classifier as it is now trained on a more balanced dataset. By having more samples of the underrepresented class, the classifier can learn better the characteristics of this class, which can lead to improved accuracy and less bias. Another identified issue is balanced accuracy, which is sometimes misleading because it does not consider the class distribution in the data. In a multi-class classification problem with imbalanced data, a model may achieve highly balanced accuracy by making correct predictions for the majority class and mostly incorrect predictions for the minority class. This can result in a misleading evaluation of the performance of the model, as it is not accurately reflected that the model can make correct predictions for the minority class, which is often the class of interest in imbalanced data scenarios. To overcome this issue, alternative evaluation metrics such as recall, F1-score, and precision are used in imbalanced data scenarios to evaluate the model's performance more accurately. This study has adopted the same metrics.

In relation to RQ1: To what extent an AI-enabled tool can be designed to generate a report for indicators and characteristics of engaging teaching videos based on teachers' movements and behaviours? The results show that a deep learning model can be trained with the indicators and characteristics of engaging teaching videos based on teachers' movements and behaviours. The AI-enabled model achieved the results with average precision, recall, f1-score, and balanced accuracy of 68%, 75%, 73%, and 79%, respectively, in classifying the annotated videos at the indicator level. The findings in this research will also assist future researchers in creating a similar AI-enabled instrument, as the authors have also provided the procedure of manual video annotation (Please refer to the attachment).

The study highlights the importance of fine-grained labelling, dataset rebalancing, and appropriate evaluation metrics in developing accurate and efficient systems for analysing teaching behaviours in educational videos. While the findings are promising, further exploration and refinement are needed, including collecting a more balanced and comprehensive dataset and incorporating additional features or modalities to enhance the system's capability (Ding et al., 2022).

In conclusion, the study focused on developing a video annotation and classification system for analysing teaching behaviours. The prototypes addressed challenges related to imbalanced datasets and misleading metrics, demonstrating improvements in performance and reliability. The findings emphasise the importance of fine-grained labelling, dataset rebalancing, and appropriate evaluation metrics in developing accurate systems for analysing teaching behaviours. Integrating AI into education has the potential to enhance online teaching effectiveness, but considerations regarding bias, privacy, job displacement, and academic integrity must be taken into account. Adjusting learning goals and leveraging AI tools can further support teaching and learning in online environments.

7.2. Implication for teaching and learning

To answer RQ2: How will the AI-enabled tool improve teaching and learning practices in higher education? This study can have at least four implications for teaching in higher education involving video conferencing. First, the AI-enabled tool developed in this study will identify the engagement enhancing teachers' behaviours and movements in the form of characteristics and indicators of engaging teaching videos and generate a report every time a recorded lecture video is processed. The AI system will efficiently analyse engaging video elements and provide recommendations, thus saving teachers considerable time recreating engaging videos. This, in turn, can yield improved student learning outcomes. The report from the AI system will highlight the timestamps of characteristics and indicators present or missing in the teaching using bar charts. These characteristics and indicators enhance students' sense of competence and connectedness, facilitate learning, and increase students' engagement. For example, if teachers encourage active participation in their video conferencing, it positively affects students' devotion to the academic experience.

Similarly, when teachers demonstrate empathy in their sessions, students feel motivated to strive for academic achievement. On the other hand, the absence of these characteristics and indicators can disengage the students. For example, if establishing teacher presence is missing in video conferencing, it can disengage the students as they develop feelings of isolation. AI reports identifying engaging video characteristics also offer teachers valuable assistance in multiple ways, including time-saving, enhanced learning, professional development support, and fostering continuous teaching improvement. Regarding professional development and ongoing improvement, AI reports help teachers identify both the strengths and weaknesses of their videos in terms of engagement. Analysing which videos successfully incorporate engaging elements provides insights for enhancing video content and innovatively refining teaching practices. Similarly, by assessing the most engaging videos, teachers gain valuable insights into what resonates best with their students, enabling informed decisions for future learning and better outcomes. Therefore, this identification of engagement enhancing behaviours and movements can assist teachers in improving their teaching and making their video conferencing more engaging for students.

Second, this AI-enabled tool can assist the institutes in creating moderation in their lecture delivery. To create moderation, the institutes can refer to the characteristics and indicators the AI-enabled tool can identify. The report generated by the AI-enabled tool can be compared for the teachers to ensure the lecture delivery is aligned with the standards and is fair, valid, consistent and reliable. Thirdly, this tool can identify the effective use of technology during video conferencing. Using technology while video conferencing, such as enabling class recording for later review, sharing screen, enabling chat function, using a microphone and camera, and varying the presentation media can increase student satisfaction levels. Lastly, this research highlights AI's importance in identifying the gaps and increasing student engagement and provides suggestions and procedures to create similar AI tools that can improve teaching. The attached manual video annotation procedure can assist future researchers in developing similar tools to improve learning and teaching.

7.3. Future research

In the last phase of this research, the proposed methodology will undergo evaluation and comparison with the understanding of educational experts. This evaluation aims to ensure that the tool is free from internal bias. To accomplish this, a new set of videos will be annotated and classified using the model presented in this study, with the resulting data being recorded for later analysis. The same videos will then be provided to educational experts for manual classification. Subsequently, the two sets of results will be compared to assess the accuracy and effectiveness of the artificial intelligence model in video classification. After the validation process, the authors will make this tool available for educational institutes, where institutes can use this to create moderation in their lecture delivery via video conferencing. This tool will also be made available for teachers who would like to identify their behaviours and movements that are present and those who are missing while presenting their lectures on video conferencing tools.

8. Limitations

The recorded lecture videos used in training the AI-enabled tool have English language teachers; for that reason, the tool may not be able to identify the indicators and characteristics of engaging teaching in other language videos. The authors trained the deep learning model with only 25 recorded lecture videos; therefore, the deep learning model is not trained with all the characteristics and indicators identified in phase 1.

9. Conclusion

This study adds to the existing knowledge of engaging teaching videos through the development of an AI-enabled tool. The tool generates a report by identifying the indicators and characteristics of engaging teaching videos where the importance of teachers' movements and behaviours in fostering student engagement has already been established in phase 1. With the help of this tool, educational institutes can create moderation in online lecture delivery. The procedure to build an AI-enabled tool can assist future researchers in creating similar tools to improve student engagement in online learning. The teachers can utilise the report from this tool to identify the gap in their teaching, and by addressing those gaps, they can make their teaching more effective and engaging.

Statements on open data and ethics

This research was carried out with ethical guidelines in mind. 25 recorded lecture videos were annotated to train the AI-enabled tool from a larger preexisting data set under ethical approval. The videos were from teachers from the University of Southern Queensland (a regional university of Australia) presenting the lecture to students online using Zoom (a video conferencing tool). The authors have ethics approval from the University of Southern Queensland with ethics approval number H20REA185.

Submission declaration

This paper has not been published previously and is not under consideration for publication elsewhere.

Funding source declaration

This research did not receive any specific grant from public, commercial, or not-for-profit funding agencies.

Author contributions

Navdeep Verma: Conceptualization, Methodology, Formal analysis, Writing - Original Draft and Review and editing.
Seyum Getenet: Conceptualization, Writing - Original Draft and Review and editing.
Christopher Dann: Conceptualization, Writing - Original Draft and Review and editing.
Thanveer Shaik: AI methodology, Formal analysis, Review and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.caeai.2023.100187.

Acronyms

DBR	Design Based Research
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
MOOC	Massive Open Online Course
RNN	Recurrent Neural Networks
CNN	Convolutional Neural Networks
LSTM	Long short-term memory
LR	Logistc Regression
ANN	Artificial Neutral Networks
SVM	Support Vector Machine
DNN	Deep Neutral Networks
VIA	VGG Image Annotator
3-D	Three-Dimensional
RGB	Red Blue Green
2-D	Two-Dimensional
AUC	Area Under the Curve
KNN	K-Nearest Neighbors

References

Aelterman, N., Vansteenkiste, M., Haerens, L., Soenens, B., Fontaine, J. R. J., & Reeve, J. (2019). Toward an integrative and fine-grained insight in motivating and demotivating teaching styles: The merits of a circumplex approach. *Journal of Educational Psychology, 111*(3), 497–521. https://doi.org/10.1037/edu0000293

Al-Samarraie, H. (2019). A scoping review of videoconferencing systems in higher education. *International Review of Research in Open and Distance Learning, 20*(3). https://doi.org/10.19173/irrodl.v20i4.4037

Aladsani, H. K. (2021). University students' use and perceptions of telegram to promote effective educational interactions: A qualitative study. *International Journal of Emerging Technologies in Learning (Ijet), 16*(9), 182. https://doi.org/10.3991/ijet.v16i09.19281

Ali, H., Qadir, J., & Shah, Z. (2023). *Chatgpt and large language models (LLMS) in healthcare: Opportunities and risks*. https://doi.org/10.36227/techrxiv.22579852.v2

Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational Researcher, 41*(1), 16–25. https://doi.org/10.3102/0013189X11428813

Authors. (2023). *Research and practice in technology enhanced learning*.

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *The Journal of the Learning Sciences, 13*(1), 1–14. https://doi.org/10.1207/s15327809jls1301_1

Beaver, I., & Mueen, A. (2022). On the care and feeding of virtual assistants: Automating conversation review with AI. *AI Magazine, 42*(4), 29–42. https://doi.org/10.1609/aaai.12024

Bergdahl, N. (2022). Engagement and disengagement in online learning. *Computers & Education, 188*, Article 104561. https://doi.org/10.1016/j.compedu.2022.104561

Bhatti, Y. K., Jamil, A., Nida, N., Yousaf, M. H., Viriri, S., & Velastin, S. A. (2021). Facial expression recognition of instructor using deep features and extreme learning machine. *Computational Intelligence and Neuroscience*, 1–17. https://doi.org/10.1155/2021/5570870, 2021.

Bourguet, M.-L., Jin, Y., Shi, Y., Chen, Y., Rincon-Ardila, L., & Venture, G. (2020). Social robots that can sense and improve student engagement. In *2020 IEEE international conference on teaching, assessment, and learning for engineering (TALE)*. https://doi.org/10.1109/tale48869.2020.9368438

Burke, K., Fanshawe, M., & Tualaulelei, E. (2022). We can't always measure what matters: Revealing opportunities to enhance online student engagement through pedagogical care. *Journal of Further and Higher Education, 46*(3), 287–300. https://doi.org/10.1080/0309877X.2021.1909712

Cents-Boonstra, M., Lichtwarck-Aschoff, A., Lara, M. M., & Denessen, E. (2021). Patterns of motivating teaching behaviour and student engagement: A microanalytic approach. *European Journal of Psychology of Education, 5*(1), 1–5. https://doi.org/10.1007/s10212-021-00543-3

Cesari, V., Galgani, B., Gemignani, A., & Menicucci, D. (2021). Enhancing qualities of consciousness during online learning via multisensory interactions. *Behavioral Sciences, 11*(5), 57. https://doi.org/10.3390/bs11050057

Cotton, D., Cotton, P., & Shipway, J. R. (2023). *Chatting and cheating. Ensuring academic integrity in the era of chatgpt*. https://doi.org/10.35542/osf.io/mrz8h

Cucinotta, D., & Vanelli, M. (2020). WHO declares COVID-19 a pandemic. *Acta BioMedica: Atenei parmensis, 91*(1), 157. https://doi.org/10.23750/abm.v91i1.9397

De Meester, A., van Duyse, F., Aelterman, N., de Muynck, G. J., & Haerens, L. (2020). An experimental, video-based investigation into the motivating impact of choice and positive feedback among students with different motor competence levels. *Physical Education and Sport Pedagogy, 25*(4), 361–378. https://doi.org/10.1080/17408989.2020.1725456

Dembereldorj, Z. (2021). Exploring online student engagement during the COVID-19 pandemic in Mongolia. *International Journal of Higher Education, 10*(7), 10. https://doi.org/10.5430/ijhe.v10n7p10

Dewan, M. A. A., Murshed, M., & Lin, F. (2019). Engagement detection in online learning: A review. *Smart Learning Environments, 6*(1), 2–8. https://doi.org/10.1186/s40561-018-0080-z

Ding, A. E., Glazewski, K., & Pawan, F. (2022). Language teachers and multimodal instructional reflections during video-based online learning tasks. *Technology, Pedagogy and Education, 31*(3), 293–312. https://doi.org/10.1080/1475939x.2022.2030790

Döring, N., Moor, K. D., Fiedler, M., Schoenenberg, K., & Raake, A. (2022). Videoconference fatigue: A conceptual analysis. *International Journal of Environmental Research and Public Health, 19*(4), 2061. https://doi.org/10.3390/ijerph19042061

Dwivedi, Y. K., Hughes, D. L., Coombs, C., Constantiou, I., Duan, Y., Edwards, J. S., … Upadhyay, N. (2020). Impact of COVID-19 pandemic on information management research and practice: Transforming education, work and life. *International Journal of Information Management, 55*, Article 102211. https://doi.org/10.1016/j.ijinfomgt.2020.102211

García-Morales, V. J., Garrido-Moreno, A., & Martín-Rojas, R. (2021). The transformation of higher education after the COVID disruption: Emerging challenges in an online learning scenario. *Frontiers in Psychology, 12*, Article 616059. https://doi.org/10.3389/fpsyg.2021.616059

Gholamrezaie, F., Bahman-Abadi, M., & Ghaznavi-Ghoushchi, M. B. (2022). *AI annotated recommendations in an efficient visual learning environment emphasising YouTube (AI-EVL)*. Cornell University - ArXiv.

Gillett-Swan, J. (2017). The challenges of online learning: Supporting and engaging the isolated learner. *Journal of Learning Design, 10*(1), 20–30. https://doi-org.ezproxy.usq.edu.au/10.5204/jld.v9i3.293.

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and CHATGPT in educational settings. *Education Sciences, 13*(7), 692. https://doi.org/10.3390/educsci13070692

Haerens, L., Aelterman, N., van den Berghe, L., de Meyer, J., Soenens, B., & Vansteenkiste, M. (2013). Observing physical education teachers' need-supportive interactions in classroom settings. *Journal of Sport & Exercise Psychology, 35*(1), 3–17. https://doi.org/10.1123/jsep.35.1.3

Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology, 15*(2). https://doi.org/10.30935/cedtech/13036

Halverson, L. R., & Graham, C. R. (2019). Learner engagement in blended learning environments: A conceptual framework. *Online Learning, 23*(2). https://doi.org/10.24059/olj.v23i2.1481

Hasnine, M. N., Bui, H. T. T., Thu Tran, T. T., Nguyen, H. T., Akçapınar, G., & Ueda, H. (2021). Students' emotion extraction and visualisation for engagement detection in online learning. *Procedia Computer Science, 192*, 3423–3431. https://doi.org/10.1016/j.procs.2021.09.115

Hew, K. F. (2016). Promoting engagement in online courses: What strategies can we learn from three highly rated MOOCS. *British Journal of Educational Technology, 47*(2), 320–341. https://doi-org.ezproxy.usq.edu.au/10.1111/bjet.12235.

Hieu, L. T., Hương, D. T., Huy, D. T. N., Dung, M. N. T. P., & Trung, N. D. (2021). Identifying learners' behavior from videos affects teaching methods of lecturers in universities. *Design Engineering*, 11146–11157. https://dinhtranngochuy.com/4117-Article%20Text-7142-1-10-20210907.pdf.

Hu, M., & Li, H. (2017). *Student engagement in online learning: A review. 2017 international Symposium on educational technology (ISET)*. https://doi.org/10.1109/iset.2017.17

Jeong, Y.-S., & Cho, N.-W. (2022). Evaluation of e-learners' concentration using recurrent neural networks. *The Journal of Supercomputing, 79*(4), 4146–4163. https://doi.org/10.1007/s11227-022-04804-w

Jia, C., Hew, K. F., Bai, S., & Huang, W. (2021). Adaptation of a conventional flipped course to an online flipped format during the Covid-19 pandemic: Student learning performance and engagement. *Journal of Research on Technology in Education, 54*(2), 281–301. https://doi.org/10.1080/15391523.2020.1847220

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data, 6*(1), 1–54. https://doi.org/10.1186/s40537-019-0192-5

Kahu, E. R., Picton, C., & Nelson, K. (2019). Pathways to engagement: A longitudinal study of the first-year student experience in the educational interface. *Higher Education, 79*(4), 657–673. https://doi.org/10.1007/s10734-019-00429-w

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Krishnamurthy, S. (2020). The future of business education: A commentary in the shadow of the covid-19 pandemic. *Journal of Business Research, 117*, 1–5. https://doi.org/10.1016/j.jbusres.2020.05.034

Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education, 79*(5), 540–563. https://doi.org/10.1080/00221546.2008.11772116

Kumar, C. B., Potnis, A., & Gupta, S. (2015). Video conferencing system for distance education. In *2015 IEEE UP section conference on electrical computer and electronics (UPCON)*. https://doi.org/10.1109/upcon.2015.7456682

Lee, J., Song, H., & Hong, A. J. (2019). Exploring factors, and indicators for measuring students' sustainable engagement in e-Learning. *Sustainability, 11*(4), 985. https://doi.org/10.3390/su11040985

Leisner, D., Zahn, C., Ruf, A., & Cattaneo, A. (2020). Different ways of interacting with videos during learning in secondary physics lessons. *Communications in Computer and Information Science*, 284–291. https://doi.org/10.1007/978-3-030-50729-9_40

Lieux, M., Sabottke, C., Schachner, E. R., Pirtle, C., Danrad, R., & Spieler, B. (2021). Online conferencing software in radiology: Recent trends and utility. *Clinical Imaging, 76*, 116–122. https://doi.org/10.1016/j.clinimag.2021.02.008

Liu, Z., Kong, X., Chen, H., Liu, S., & Yang, Z. (2023). MOOC-BERT: Automatically identifying learner cognitive presence from MOOC discussion data. *IEEE Transactions on Learning Technologies, 1*. https://doi.org/10.1109/tlt.2023.3240715. –14.

Liu, Y., Ma, J., Xie, Y., Yang, X., Tao, X., Peng, L., & Gao, W. (2022). Contrastive predictive coding with transformer for video representation learning. *Neurocomputing, 482*, 154–162. https://doi.org/10.1016/j.neucom.2021.11.031

McKenney, S., & Reeves, T. (2018). *Conducting educational design research* (2nd ed.). Routledge. https://doi.org/10.4324/9781315105642

Miah, S. J., Solomonides, I., & Gammack, J. G. (2020). A design-based research approach for developing data-focussed business curricula. *Education and Information Technologies, 25*(1), 553–581. https://doi.org/10.1007/s10639-019-09981-5

Mishra, L., Gupta, T., & Shree, A. (2020). Online teaching-learning in higher education during lockdown period of COVID-19 pandemic. *International Journal of Educational Research Open, 1*, Article 100012. https://doi.org/10.1016/j.ijedro.2020.100012

Mubarak, A. A., Cao, H., & Ahmed, S. A. M. (2020a). Predictive learning analytics using deep learning model in MOOCs' courses videos. *Education and Information Technologies, 26*(1), 371–392. https://doi.org/10.1007/s10639-020-10273-6

Mubarak, A. A., Cao, H., Zhang, W., & Zhang, W. (2020b). Visual analytics of video-clickstream data and prediction of learners' performance using deep learning models in MOOCs' courses. *Computer Applications in Engineering Education, 29*(4), 710–732. https://doi.org/10.1002/cae.22328

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality, 15*(2), 1–21. https://doi.org/10.1145/3597307

Oh, E., & Reeves, T. C. (2010). The implications of the differences between design research and instructional systems design for educational technology researchers and practitioners. *Educational Media International, 47*(4), 263–275. https://doi.org/10.1080/09523987.2010.535326

Pabba, C., & Kumar, P. (2021). An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Systems, 39*(1). https://doi.org/10.1111/exsy.12839

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: Chatgpt and beyond. *Journal of University Teaching and Learning Practice, 20*(2). https://doi.org/10.53761/1.20.02.07

Roth, J. J., Pierce, M., & Brewer, S. (2020). Performance and satisfaction of resident and distance students in videoconference courses. *Journal of Criminal Justice Education, 31*(2), 296–310. https://doi.org/10.1080/10511253.2020.1726423

Salas-Pilco, S. Z., Yang, Y., & Zhang, Z. (2022). Student engagement in online learning in Latin American higher education during the COVID-19 pandemic: A systematic review. *British Journal of Educational Technology, 53*(3), 593–619. https://doi.org/10.1111/bjet.13190

Shaik, T., Tao, X., Higgins, N., Gururajan, R., Li, Y., Zhou, X., & Acharya, U. R. (2022). FedStack: Personalised activity monitoring using stacked federated learning. *Knowledge-Based Systems, 257*, Article 109929. https://doi.org/10.1016/j.knosys.2022.109929

Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J., Redmond, P., & Galligan, L. (2022). A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *IEEE Access, 10*, 56720–56739. https://doi.org/10.1109/access.2022.3177752

Sharma, P., Joshi, S., Gautam, S., Maharjan, S., Khanal, S. R., Reis, M. C., Barroso, J., & de Jesus Filipe, V. M. (2022). Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. *Communications in Computer and Information Science*, 52–68. https://doi.org/10.1007/978-3-031-22918-3_5

Singh, V., & Thurman, A. C. (2019). How many ways can we define online learning? A systematic literature review of definitions of online learning (1988-2018). *American Journal of Distance Education, 33*(4), 289–306. https://doi.org/10.1080/08923647.2019.1663082

Tao, X., Shaik, T. B., Higgins, N., Gururajan, R., & Zhou, X. (2021). Remote patient monitoring using radio frequency identification (RFID) technology and machine learning for early detection of suicidal behaviour in mental health facilities. *Sensors, 21*(3), 776. https://doi.org/10.3390/s21030776

Thomas, C., & Jayagopi, D. B. (2017). Predicting student engagement in classrooms using facial behavioural cues. In *Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education*. https://doi.org/10.1145/3139513.3139514

Torrato, J. B., Aguja, S. E., & Prudente, M. S. (2021). Using web video conferencing to conduct a program as a proposed model toward teacher leadership and academic vitality in the Philippines. *Education Sciences, 11*(658), 1–30. https://doi.org/10.3390/educsci11110658

Trenholm, S., Hajek, B., Robinson, C., Chinnappan, M., Albrecht, A., & Ashman, H. (2019). Investigating undergraduate mathematics learners' cognitive engagement with recorded lecture videos. *International Journal of Mathematical Education in Science & Technology, 50*(1), 3–24. https://doi.org/10.1080/0020739x.2018.1458339

Van, D. A. J., Gravemeijer, K., McKenney, S., & Nieveen, N. (Eds.). (2006). *Educational design research*. Taylor & Francis Group. https://doi.org/10.4324/9780203088364.

Vrskova, R., Hudec, R., Sykora, P., Kamencay, P., & Radilova, M. (2020). *Education of video classification based by neural networks. 2020 18th international Conference on Emerging elearning Technologies and applications (ICETA)*. https://doi.org/10.1109/iceta51985.2020.9379190

Wang, Q., Huang, C., & Quek, C. L. (2018). Students' perspectives on the design and implementation of a blended synchronous learning environment. *Australasian Journal of Educational Technology, 34*(1), 1–13. https://doi.org/10.14742/ajet.3404

Wróblewska, A., Jasek, J., Jastrzebski, B., Pawlak, S., Grzywacz, A., Cheong, S. A., & Hołyst, J. (2022). Deep learning for automatic detection of qualitative features of lecturing. *Lecture Notes in Computer Science*, 698–703. https://doi.org/10.1007/978-3-031-11644-5_70

Zaletelj, J., & Košir, A. (2017). Predicting students' attention in the classroom from Kinect facial and body features. *EURASIP Journal on Image and Video Processing, 2017*(1). https://doi.org/10.1186/s13640-017-0228-8

Zhai, X. (2022). *ChatGPT user experience: Implications for education*. https://doi.org/10.2139/ssrn.4312418. SSRN 4312418.