Education and Information Technologies
https://doi.org/10.1007/s10639-024-12931-5

Evaluating the impact of an automated body language assessment system

Eleni Dimitriadou[1,2] · **Andreas Lanitis**[1,2]

Received: 8 January 2024 / Accepted: 25 July 2024

Abstract

The body language of an educator during a class can affect student's level of interest and concentration. As an attempt to assist educators to improve their body language and speaking characteristics, a pilot body language analysis system that assesses the body language of educators was developed. The proposed application makes use of specific biometric features for determining body language quality during class delivery. The aim of the current study is to examine whether the proposed application can contribute to improving the teachers' body language, whether the application can provide satisfactory feedback related to the teachers' body language, and whether the use of the application in real classroom conditions is acceptable. As part of this effort the pilot application has been assessed by teachers of primary, secondary and university education. The experimental investigation involved two phases. In the first phase participants delivered a short lecture that was evaluated using the automated body language analysis application. After the lecture participants were informed about the operation of the application and they were presented with the feedback generated by the body language analysis. During the second phase participants delivered a second short lecture. By comparing the body language quality between the two phases, conclusions related to the impact of the application in improving body language were derived. Experimental results demonstrate that the application provides satisfactory feedback, it is acceptable to use the application in real class conditions, and the feedback provided can be used for self-assessment, reflection and improvement regarding educator's body language.

Keywords Automated body language analysis tool · Lecturer evaluation · Body language quality · Impact assessment

Extended author information available on the last page of the article

Published online: 21 August 2024

1 Introduction

Body language is an important aspect in educational settings, directly influencing educators' effectiveness in message delivery, classroom management, and student interactions (Benzer, 2012). Automated lecture quality evaluation tools can provide objective and fast feedback on the lecture delivery quality, supporting educators in improving their lecture presentation quality and, as a result, improving students' learning experiences. As part of an effort to assist educators to improve their body language, a Body Language Assessment (BLA) application that provides automated feedback regarding the quality of educator's body language was developed (See Fig. 1) (Dimitriadou & Lanitis, 2023). The BLA application relies on a set of measurable biometric features collected from video and audio recordings, to estimate body language quality, providing in that way an easy to use and low-cost alternative for performing a task that usually requires human expertise. The biometric characteristics used for assessing body language quality were defined through a systematic multi-phased process that includes the definition of a set of body language quality indicators reflecting the goodness of a lecture presentation. The biometric features considered relate to facial expressions, body activity, hand movements, facial pose, speech rate and speech tone which can be retrieved from video recordings of lectures in real time or in a batch processing mode. So far, few models have been proposed that combine hand and facial pose features due to the challenges that exist in tracking human faces from video (such as hand and head tilt) (Rastgoo et al., 2021). The main steps in the process of defining suitable quality metrics include the requirements analysis step, in which a set of biometric features associated with lecture delivery quality are defined through a literature review and interviews with stakeholders, followed by the development of dedicated methods for extracting the features from a video stream. The extracted features are then utilized to calculate an overall lecture delivery quality score.

Due to the importance of high-quality body language in the educational process several attempts to evaluate and provide feedback about body language were recorded in the literature. In some cases, questionnaire-based methods are used for body language evaluation. For example, Gulec and Temel (2015) explore the differences in body language usage between prospective elementary mathematics and social sciences teachers based on data gathered using a 21-item "Body Language

Fig. 1 Automated BLA tool. A video of a lecture captured by an ordinary camera is analysed to extract body language quality indicators

Questionnaire (BLQ)". Findings show a difference in body language usage between the two groups of teachers, with positive attitudes towards body language utilization in teaching. Kucuk (2023) explores teachers' perceptions of the role and importance of body language in the educational setting. The user evaluation was conducted through a questionnaire distributed to teachers and in-person interviews. The questionnaire was designed to gather insights into the teachers' opinions on the significance of body language in teaching. On the other hand, Tai (2014) explores the utilization of body language in English teaching to foster a more engaging and effective learning environment. It delves into how teachers use gestures, facial expressions, and eye contact to enhance communication and understanding in the classroom. The approach relies on observational insights and existing literature to argue the benefits of using body language in English teaching, without specifying a concrete user evaluation method. Rosen et al. (2015) promotes the "teacher-as-researcher" paradigm, encouraging sign language teachers to engage in action research to enhance teaching and learning experiences in their classrooms. User evaluation is executed through action research, a stepwise process where teachers identify gaps in the current teaching approach, develop interventions, and assess the outcomes. This hands-on approach to research is grounded in real classroom experiences, aiming to bridge the theoretical and practical aspects of teaching.

Artificial Intelligence (AI) in education has a crucial role in solving educational crises (Ouyang et al., 2022). Few systems that use automated approaches (based on AI) to assess lecture quality and/or lecturer's body language are, also reported in the literature (Mohammadreza and Safabakhsh (2021), Zhao and Tang (2019), Dimitriadou & Lanitis (2023)). Mohammadreza and Safabakhsh (2021) presents a novel approach to assessing lecture quality based on audience reactions, specifically focusing on students' engagement levels. They developed a system that analyses students' facial reactions, captured in real-time, to infer engagement. This research is significant in the educational field as it shifts from traditional assessment methods, such as questionnaires or external observers, to an automated system trained using machine learning (ML). However, the proposed system relies on videos showing students, hence the implementation of the system in real situations may not be acceptable due to data privacy issues. Furthermore, Zhao and Tang (2019) developed a classroom teaching quality evaluation model using a BP (Back Propagation) neural network algorithm. The model uses student evaluations and teaching team assessments as input data, in order to evaluate teaching quality. This model improved on existing methods by reducing subjectivity, avoiding overfitting, and increasing accuracy. Dimitriadou & Lanitis (2023), focus their attention on developing an integrated lecture quality evaluation application for lecture quality assessment, as well as performing a performance evaluation experiment for determining the accuracy of the system in estimating body language quality scores.

According to Van Klaveren (2011), "lecturing style" refers to the teaching method where the instructor primarily delivers information through traditional lecturing in front of the class. The study investigates how the proportion of time spent on this lecturing style affects student performance in mathematics and physics in the Netherlands. In the other hand, according to Saroyan and Snell (1997), "lecturing style" encompasses variations in how lectures are delivered, ranging from didactic dissemination of content

to pedagogically sophisticated approaches. The study examines different instructional plans and protocols to understand how variations in lecturing style impact student evaluations and learning outcomes. Van Klaveren (2011) and Saroyan and Snell (1997) highlight the importance of understanding the impact of lecturing style on student performance and emphasize the need to consider variations in instructional approaches for effective teaching. In contrast to the above studies, we focused our attention on evaluating the acceptance of an automated system that evaluates the quality of teachers' body language during class delivery.

While there is considerable research on body language in education and teacher evaluation systems, there is a lack of studies focusing exclusively on the automated evaluation of teachers' body language, its acceptance by educators or students, and most importantly there is lack of studies that assess the impact of such applications in improving body language. The present study aims to fill the gap in the literature concerning the acceptance and impact of such systems by the stakeholders. Evaluating the acceptance of an application that assesses body language by educators, who are directly implicated, is paramount for several reasons. Firstly, educators are the primary users and thus, their comfort, ease of use, and trust in the application will directly influence its successful implementation. Secondly, their first-hand experience in the educational setting enables educators to provide valuable feedback on the practicality and relevance of the application in real-world teaching scenarios. Lastly, ensuring that such an application is well-received and valued by educators not only enhances its credibility but also fosters a positive attitude towards technology-assisted teaching methodologies, thereby facilitating a smoother integration of such technologies in educational settings. This, in turn, can pave the way for more data-driven and evidence-based teaching strategies, enhancing the overall educational experience for both teachers and students.

In the remaining sections, we present a literature review on educator performance evaluation (Section 2) and describe the body language analysis application (Section 3). In Section 4 the experimental evaluation procedure and results of the evaluation are presented followed by a discussion and limitations (Section 5 and Section 6 respectively). Conclusions and plans for future work are presented in Section 7.

2 Literature review

A lecturer's performance is crucial for ensuring the quality of lectures delivered to students. Traditional lecturer performance evaluation has been a complex and time-consuming process that does not provide instantaneous and seamless feedback. This section provides a brief description of the importance of body language and educator performance assessment, and also presents methodologies used for assessing teacher performance during class delivery.

2.1 Educator performance assessment

To assess teaching quality in educational institutions, a range of factors concerning evaluation are frequently considered as a means of providing a comprehensive view

regarding the educational experience. These factors might involve the knowledge of learners (Dubinsky et al., 2022), subject knowledge (Schempp et al., 1998), teaching methodology (Darling-Hammond et al., 2013), general pedagogical knowledge (Liakopoulou, 2011), curriculum knowledge (Behar & George, 2013), knowledge of contexts (Darling-Hammond et al., 2010), lecture style, self-knowledge (Schussler et al., 2010) and audience interaction (Short & Martin, 2011). Practically, the effectiveness of an educator can be determined by the interrelation of those factors along with the competence to apply and integrate them completely to make optimal student development and learning easier. As far as the current study is concerned, lecture quality assessment is performed in relation to the lecturing style, rather than considering the general teaching performance of an educator which involves all aforementioned categories.

Traditional assessment methods are frequently based on teachers' observation by experts during class delivery, resulting in a process that can be costly, not accurate, time-consuming and most of the time, the provided feedback is not frequent and is associated to the actual performance and not on the way the teachers can improve their techniques (Archer et al., 2016). In some cases, student feedback is used for assessing the performance of educators. However, this method is considered an unreliable method because the data is not collected in real-time and there might be manipulative responses (Winarno, 2017). To face this significant impediment in the development of a teacher, new technologies could be employed for the production of high quality and important automatic lecture quality feedback for teachers.

2.2 Teachers' body language during class delivery

Body language refers to nonverbal cues and behaviors, communicating emotions, intentions, and thoughts. This includes facial expressions, gestures, posture, eye movements, touch, and even aspects of speech such as tone and volume (Fast, 1970; Pease, 1981). According to Kucuk (2023), body language in education is defined as the use of non-verbal signs (such as posture, facial expressions, gestures and body language) by teachers and students to convey information and express emotions during the learning process.

Body language plays a vital role in education as it can influence learning and teaching dynamics. Teachers' non-verbal cues can enhance clarity, keep students engaged, and establish a positive classroom environment. For students, their body language can reflect their level of understanding, engagement, and emotions, providing teachers with feedback. Recognizing and responding to these cues can significantly enhance the educational experience (Miller, 2005; Kucuk 2023; Hussain et al., 2022). Teachers who effectively use body language create a positive and attractive environment (Kucuk 2023) and students showing higher levels of satisfaction for the lesson (Caswell & Neill, 2003).

Teachers' body language can be evaluated either in an automated way or in a non-automated way. The use of automated methods requires technological tools (Abdulrahaman et al., 2020; Turaev et al., 2023), while the use of non-automated methods is based on observations, surveys (Tai, 2014), questionnaires (Kucuk 2023), and

interviews (Denham & Onwuegbuzie, 2013). Specifically, non-automatic measurement of body language relies on meticulous human observation, including coding behaviors from videos, gathering self-reports, and expert observation in real-time. Despite drawbacks like intensive labor and observer biases, these methods yield invaluable insights through keen attention to details.

2.3 Non-automatic methods for body language evaluation

The following studies are focused on evaluating body language using non-automated methods. Hattie and Timperley (2007) examined effective methods of feedback without specifically focusing on body language. They linked evaluation to effective teaching and explored how teachers could use feedback to improve their teaching practices. While this study does not exclusively focus on body language, it underscores the importance of effective evaluation and feedback in enhancing teaching and learning. Although several studies examine body language in the educational sphere, only few studies focus exclusively on the evaluation of teachers' body language and the acceptance of these evaluation systems by other educators or students. Woolfolk and Brooks (1985) investigated the impact of non-verbal cues, including body language, on students' perceptions of teachers. Their findings indicated that students' perceptions of teachers are significantly influenced by non-verbal cues, which, in turn, can impact their acceptance of teacher evaluation systems.

Tok and Temel (2014) deal with the quantification of teachers' body language proficiency. Within this context, they designed a 23-item scale through a rigorous process that incorporated feedback from educators, analysis of existing research, and pilot testing. When applied to a sample of 503 pre-service teachers, the scale demonstrated reliability, backed by significant statistical indicators. Building on this foundational work, Karaca and Filiz (2023) employed Tok and Temel's scale to assess the body language competencies of Physical Education Teachers (PETs), surveying a total of 347 PETs. Their findings indicated that those PETs who had received training in communication skills displayed more advanced body language proficiency. The norms/criteria explained by Tok and Temel (2014) include: (a) Assessment of body language: Describes how the evaluation of educators' body language is conducted using specific criteria or a rating scale. (b) Validity and reliability: Explain how the validity and reliability of the body language assessment scale are checked. (c) Correlation with effective teaching: Analyzes the correlations between the body language assessment scale and the effectiveness of teaching, indicating the significance of the results for the quality of education. Tok and Temel (2014) focused on communication skills and body language and the elements identified from the literature were grouped into four categories: sound and intonation, clothing, posture, gestures, and facial expressions. The present study is based on three of the four categories of the study by Tok and Temel (2014), sound and intonation, posture, gestures, and facial expressions. However, the present study also includes additional indicators that have emerged through semi-structured interviews with chief education officers, educators, and students (see Section 3).

Bower et al. (2013) undertook a study to understand how various communication elements could impact pre-service teachers' presentations. Their research highlighted the importance of alignment between body language, voice, and verbal content. To assess the presentation abilities of pre-service teachers, they contrasted two assessment models: the "Constructed Impression" model and the "Modes of Communication" model. The former focused on aspects such as confidence and clarity in communication, while the latter emphasized the integration of voice, words, and body language. Results from Bower et al. (2013) indicated that the "Constructed Impression" model provided a more accurate representation of presentation skills, but the "Modes of Communication" model was more predictive of overall performance scores. Furthermore, Kucuk's (2023) employed a mixed-method approach to gain insights into teachers' perspectives on body language. Utilizing both a questionnaire and subsequent interviews with 30 educators from Tishk International University, Kucuk's findings consistently underscored the critical role that body language plays in effective instruction.

2.4 Automated body language performance evaluation

In few cases attempts for automating the process of teacher performance evaluation were recorded. Bhatia and Kaur (2021) employ IoT systems in their classes to gather information concerning the students' and educators' performance so as to recognize their progress. The data collected from students is associated with various activities of students performed while being at school (e.g., teamwork, attendance, academic performance) whereas teachers' data is associated with their performance evaluation (e.g., number of assignments, student satisfaction, quality of content). By using the Bayesian modelling approach, the information gathered is evaluated via a device of fog-cloud computing with the purpose to detect a quantifiable performance measure. In addition, this measure of progress is calculated over time for assessing educators' and students' performance. The outcome of the specific method is noticed through the experiments which are conducted by employing four datasets and indicate the method's efficiency.

Yang et al. (2012) suggested a teaching assessment model to face the complexities demonstrated in evaluating teaching quality. In the proposed teaching model, they utilize back propagation networks to calculate the teaching evaluation index concept quantitatively. The characteristics employed for teaching quality evaluation are connected to guiding ideology, teaching conditions, teachers, construction of study style, teaching construction and reform, teaching effectiveness, teaching management, classroom teaching effectiveness, and others. The reported results proved that the current model has a wide applicability in evaluating the effectiveness of teaching. Despite the fact that the research illustrates positive results, several elements of their approach justify further scrutiny. One limitation worth mentioning is the lack of metrics of comprehensive performance to assess the effectiveness of the model in an objective way. While the authors state that there were satisfactory results and the neural network outputs align closely with the target values, the absence of certain metrics like accuracy, recall, mean squared error or precision

weakens the quantitative assessment of the performance of the model. Hence, this study would be enhanced from an analysis that compares it with other fixed methods of teaching quality assessment or benchmarks to ensure its unique contributions and the superiority of the model.

Jensen et al. (2020) proposed a method applied to teachers as a means of effortlessly audiotaping the classroom conversations and lectures. They also employed machine learning algorithms and voice recognition as an attempt to offer generalized calculations. These estimations were given in the form of scores, which were extracted from computers, and they were considered as crucial parts of educator speech. In particular, the authors evaluated the audio quality from the recording of teachers with rates of A, B, C or F. “A” indicated an outstanding quality of the recording, "B" demonstrated a satisfactory quality with minor volume or background noise issues, recordings denoted with a "C" had flawed segments, and an "F" indicated that the audio files included irreparable technical errors or were lost. By comparing them with human interpreters, they noticed that automatic methods were almost precise and that the mistakes of voice recognition had almost no effect on performance. Thus, they argue that the conversation of the actual instructor can be recorded and assessed providing automated feedback. Moreover, the automated algorithms might be used into/ a dynamic visualization system offering teachers the required feedback regarding their level of speech. The comparison indicates the assessment of the actual speaker performance since these particular algorithms use machine learning and voice recognition to offer overall estimations and scores that were given by computers, specifically evaluating significant parts of educator’s speech. This proves that the emphasis is given on the evaluation of the content and also on the delivery of the instructor’s speech instead of concentrating on the audio recording technical quality. Additionally, the feedback provided to teachers is shaped according to their “level of speech,” suggesting that it is relevant to calculating how the educators effectively communicate and get the students’ attention of during conversations and lectures in the classroom. On the whole, the current algorithms try to offer crucial feedback on how well the teachers give their lectures and engage with their students, by aiding them to increase their instructional effectiveness and communication skills. However, the approach used by Jensen, et al. (2020) focus solely on audio features, and consequently, it does not provide visual features associated with a lecturer’s in-class activity.

Jensen et al. (2021) focus on designing an automatic educator feedback framework which requires various considerations as for audio data processes, automatic assessment and how feedback is demonstrated. The authors use machine learning techniques, including Random Forest classifiers, with transfer learning from the Bidirectional Encoder Representations from Transformers (BERT) algorithm regarding natural language processing (NLP). BERT, a pre-trained language model based on transformer architecture, is employed to attain bidirectional contextual understanding of language, making it possible to capture contextual relationships that are intricate within sentences. The input to both the Random Forest Classifier (RF) and the Bidirectional Encoder Representations from Transformers (BERT) models includes transcribed utterances. The outcomes illustrate that BERT provides more accurate and superior input in

various degrees, making it by far the most practical technique in the provision of automated feedback on teacher discourse, exceeding the performance of several machine learning techniques, like Random Forest classifiers.

Zhu (2022) adopt the use of both Analytic Hierarchy Process (AHP) and fuzzy decision tree algorithms to assess teaching quality, particularly in the English education context. This system consists of various assessment indexes as well as procedures as an attempt to comprehensively assess the English instruction quality. The four assessment indexes, which are considered to be critical, involve Teaching Attitudes, assessing the approach and disposition of teachers; Teaching Contents, giving emphasis on the themes and subjects on the curriculum; Teaching Techniques inspecting the methodologies used during instruction; and Teaching Effects, providing an analysis of the overall efficacy and success of the experience in education. By combining methods like student evaluation, peer evaluation, expert evaluation, and self-evaluation, the study includes these specific aspects via a decision tree model. The findings enhance the improvement of the quality of teaching in colleges by offering reliable and objective evaluation results. Still, the current study is only applicable to the assessment of English lessons and therefore, it cannot be considered as a general lecture of quality evaluation method, since it was not designed or tested for subjects that are beyond the English language scope of instruction. Its criteria and methodologies might not effectively apply to other linguistic contexts or disciplines, eliminating its wider applicability.

Barmaki and Hughes (2018) developed an automated teacher body language evaluation system designed to detect and provide feedback on non-verbal "closed" gestures prevalent in virtual teaching sessions. Utilizing the kinect motion-capture device they created a gesture recognition method that analyzed full-body data. After training the system using annotated clips of five students, their Visual Gesture Builder (VGB) was employed to produce multiple classifiers for each gesture, culminating in an ensemble with a remarkable accuracy of $96.5\% \pm 2.1\%$. When tested in a case study, the system showed significant potential in enhancing the body language mindfulness of users through its visual feedback mechanism, with a follow-up study further indicating a positive reception to an added vibration feedback feature. In relation to our proposed system, Barmaki and Hughes (2018) required specialized equipment to evaluate teachers' body language (such as Kinect). Furthermore, their study body language assessment method is based on gestures rather than a multitude of biometric features such as speech recognition.

The evaluation and acceptance of teachers' evaluation systems are critical for enhancing the quality of education. While there is a vast body of literature on various aspects of teacher evaluation, there is limited research specifically focusing on the impact of such applications to improving body language. In few occasions impact assessment with multi-phased experiments was performed (Barmaki & Hughes, 2018). The work presented in this paper aims to address the issues of impact and acceptance related to automated body language quality determination.

3 Body language evaluation application

The experimental evaluation conducted in our work uses a body language assessment application developed by Dimitriadou & Lanitis (2023). The application utilizes a combination of multiple biometric features separated into five different modalities (see Table 1), as a means of estimating body language quality from a video stream showing an educator delivering a class.

The identification of indicators of low/high-quality body language was based on the existing literature as well as through semi-structured interviews with teachers, chief education officers, and students. Initially, through the literature review, an initial set of characteristics related to a lecturer's body language was defined (Azer, 2005; Bambaeeroo & Shokrpour, 2017; Tok & Temel, 2014; Mohammadreza & Safabakhsh, 2021). The findings from the relevant literature were cross verified based on the outcomes of semi-structured interviews with relevant stakeholders. During the interviews, chief education officers, educators, and students watched "low/high" lecture quality videos available on YouTube and then indicated specific actions they considered as low/high quality body language. Based on the results of the interviews and the literature findings, the indicators of low/high-quality body language have been identified (see Table 1).

During the system operation expression recognition is carried out to categorize the seven basic facial expressions (anger, disgust, fear, happy, neutral, sad, and surprise) (Lasri et al., 2019). Furthermore, this system identifies the following seven desirable and undesirable activities of educators in class: being absent from the camera's view, raising hand(s), attending, writing, texting, making a telephone call, and looking elsewhere. Regarding the speech metric, "Word Density", Speaking speed and Speaking intonation are estimated. All three speech related features are combined to provide an overall speech quality score. Hand Movement detection and the location of detected hands is considered for calculating the speed of hand movements. Hand speed is estimated as the distance between the previous and the current hand center divided by the elapsed frames number.

Table 1 Body language quality metrics

Biometric Features	High Quality Body Language	Low Quality Body Language
Facial Expressions	Happy, Surprise, Neutral	Anger, Fear, Disgust, Sad
Body Activity	Attending, Writing, Hand Raising	Absent, Telephone Call, Texting, Looking Elsewhere
Speech	Word Density (35%-55%) Speaking Speed (150–250 words per minute) Speech Intonation (40%-60%)	Word Density ($<35\%, >55\%$) Speaking Speed ($<150, >250$ words per minute) Speech Intonation ($<40\%, >60\%$)
Hand Movement	Moving	Stationary
Facial Pose	Left, Right, Up, Down and Forward	Far-Left, Far-Right, Far-Up, Far-Down, Backwards

A face detector is utilized for deciding whether the lecturer looks *left*, *right, up, down, far-left, far-right, far-up, far-down, forward and backwards.*

Speaking Speed Previous studies, state that a rate of 150–190 w.p.m (words per minute) consider as the normal or average speed (Marslen-Wilson, 1973; Richards, 1983). Recently, Wang (2021) analysed academic lectures from three sources (BBC iPlayer, YouTube, daily life's conversations) to define the average number of words per minutes and the results showed that the average w.p.m is 125–247. Based on those results the optimum w.p.m was initially set to 150–250 w.p.m. This range was cross verified as acceptable during interviews with the main stakeholders in the field of education (educators, students and chief education officers). Hence, when the speaking speed is within the predefined range it is considered a high-quality lecture style, otherwise it is associated with low quality lecture style.

Word density In the literature, it is stated that the average word density percentage for teachers is 35%-55% (Hunter & Titze 2010). This finding came into agreement with the views expressed during the interviews with the stakeholders in this field. Thus, word density values between 35%-55% are associated with high quality lecture quality, whereas values outside the recommended range are associated with low quality lecture quality.

Speech intonation According to Albergaria-Almeida (2010), teachers are expected to spend approximately 50% of their teaching time on questions. A study by the Teaching Excellence in Adult Literacy (TEAL) Center found that the teachers are spending 35–50% of their lecture time on questions to students (Corley & Rauscher, 2013). Based on the value suggested by Albergaria-Almeida (2010), and in agreement with the views expressed during the interviews, speech intonation in the range of 40%-60%, was associated with high quality lecture style and otherwise it was associated with low quality lecture style.

Facial pose According to previous studies, one of the key elements of an effective lecture presentation is for the presenter to maintain eye contact with the audience (Gelula, 1997) as this improves the interaction between a presenter and an audience (Dolan, 2017). To implement this feature, the face bounding box and the nose location, in an image frame, are used for determining ten facial poses (Forward, Left, Far-Left, Right, Far-Right, Up, Far-Up, Down, Far-Down, and looking Backwards). Figure 2 shows indicative examples of facial poses considered. Assuming that the lecturer stands in front of a class, it is determined that optical contact is maintained when the facial poses Forward, Left, Right, Up, and Down are adopted. In contrast, the facial poses Far-Right, Far-Left, Far-Up, Far-Down and Backwards do not guarantee optical contact with the audience. Participants at the interviews cross verified these observations.

Based on the feature extraction process the body language quality (between 0 and 1) is given for each of the five modalities. The accumulation of the five modality scores provides a total body language quality score in the range of zero to five

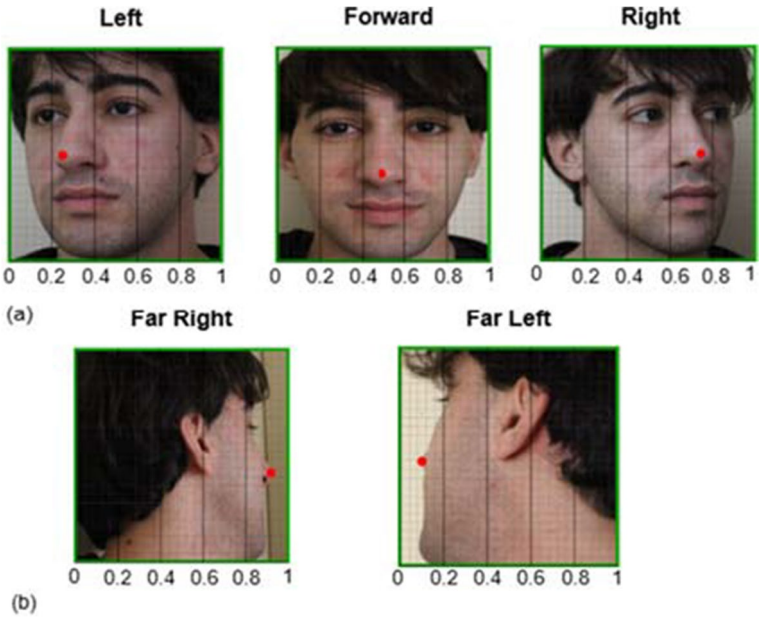

Fig. 2 Indicative facial poses associated with high (top row) and low-quality lecture style (bottom row)

for each image frame. The score is estimated in real time and can be displayed on the screen of the portable computer used for running the application, whereas as dedicated graphs related to the total score and scores per modality are presented to the educator at the end of a lecture, offering in that way the necessary feedback for reflection and self-improvement. Examples of the feedback provided by the system is shown in Fig. 3.

To assess the ability of the application to provide reasonable body language quality scores, annotated data was employed as the ground truth. To generate ground truth data a series of videos showing lectures were presented to nine annotators who provided body language quality scores for each video frame. The average scores among all volunteers for the overall body language quality and the average scores for each modality per frame are considered as the ground truth. During the evaluation process, the same videos were presented to the machine and auto generated scores were compared with annotator scores. The results of the evaluation indicate that the automated body language assessment system estimated body language quality with adequate accuracy. Furthermore, the results indicate that the performance of the system is either the same or even better when compared to individual human annotator's performance. More details about the system and the quantitative performance evaluation are presented by Dimitriadou & Lanitis (2023).

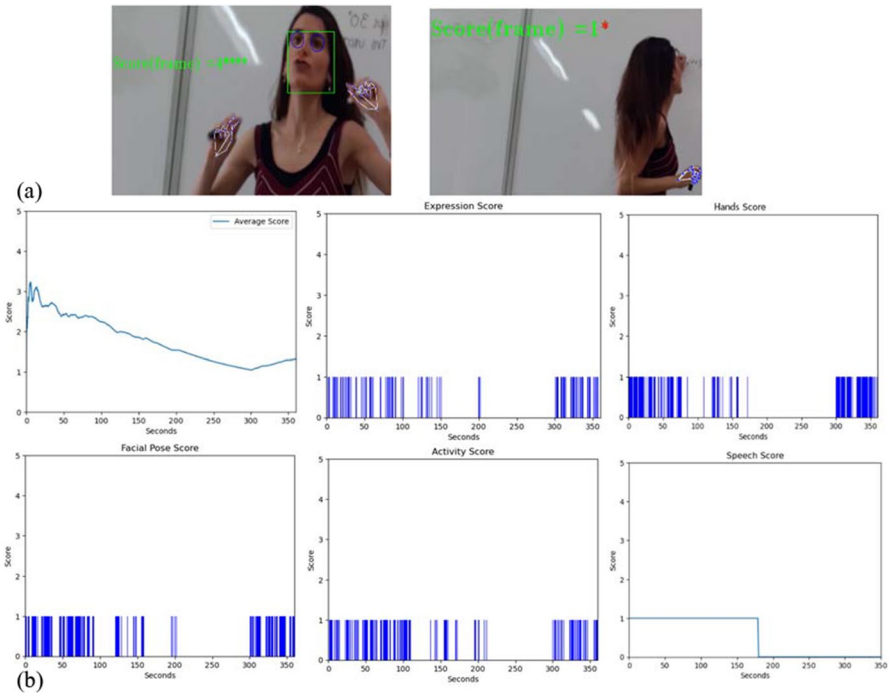

Fig. 3 Sample Feedback Provided by the System. **a** Typical screenshots provided by the system showing the instantaneous lecture quality score. **b** Typical graph of performance over whole lecture

4 Methodology

The impact of the body language assessment application is evaluated through a dedicated experimental investigation that involves comparative assessment of the impact of the application in improving body language quality. The steps adopted through the experimental evaluation, and the results obtained are described in this section.

4.1 Research questions

The purpose of this research is to examine the impact of the body language analysis experimental application in improving the body language of educators. As part of the process, the following research questions were considered:

RQ1: Can this application potentially contribute to the improvement of the educators' body language?
RQ2: Does the application provide sufficient feedback regarding the body language of educators?
RQ3: Is the use of this application acceptable in actual classroom conditions?

4.2 Overview of experimental procedure

As part of the experimental procedure, a system was developed that combines a variety of biometric features related to low/high body language quality. Biometric characteristics as well as indicators related to low/high-quality body language have been derived from the existing literature review and through semi-structured interviews with teachers, chief education officers and students (see Section 3). The experimental procedure consisted of two phases (see Fig. 4). In the first phase of the process, participants completed the first part of a questionnaire that contained closed type questions related to their demographic data such as ethnicity, gender, profile, age, field of expertise and years of teaching experience. Afterwards, the participants had the chance to present a short demo lecture and the body language analysis application (see Section 3) was used for analysing and estimating scores reflecting their body language during the lecture. After completing the first lecture, participants completed the second part of the questionnaire that included closed type questions regarding the user experience of the participants in using the BLA application. Following the completion of the second part of the questionnaire, short semi—structured interviews of the participants took place which had a duration of 10—15 min. The interviews were conducted by the researchers and contained information and a discussion related to the interpretation of results produced by the automated body language analysis application.

In the second phase of the experimental process, participants delivered another short demo lecture of the same duration as in the first phase. Then, they proceeded to the completion of the third part of the questionnaire that contained the same closed type questions as in the second part. At the end of the third part of the questionnaire the participants were given open type questions in order to provide additional feedback regarding the body language analysis application. At the end of the third part of the questionnaire a second short semi—structured interview of the participants took place. The interviews conducted by the researchers contained comments

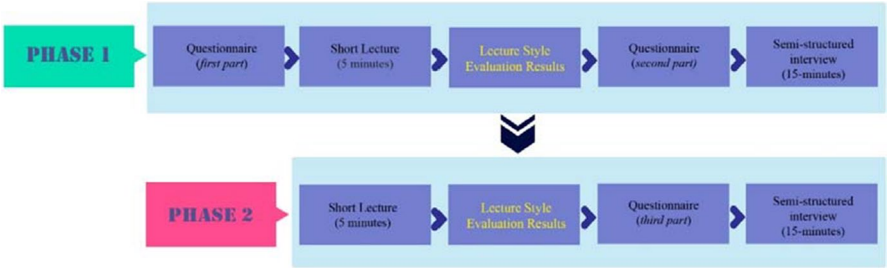

Fig. 4 Overview of the experimental procedure

of the results of the application in order to determine whether at that point there was improvement concerning the comprehension of these results in comparison to the first phase of the research.

Before the beginning of the study, the participants were informed about the purpose of the study in order to guarantee that everyone follows the same steps towards its completion. Concerning ethics, all the educators were acquainted about the purpose of this study both orally and in writing in order to obtain their informed consent. Furthermore, the participants were assured of the confidentiality and anonymity of their personal data. Educators were informed that they could withdraw from the procedure at any time without giving any explanations. The study underwent an ethics review process and was approved by the National Bioethics Committee, ensuring that the experimental investigation was conducted in accordance with ethical standards.

4.3 Sample

The sample of the main research contained nine educators of elementary, secondary, and higher education, from the age of 25 to 59. Table 2 presents a summary of the demographic characteristics of nine participants of current research. Considering experience in artificial intelligence technologies, the *22%* ($N=2$) are highly experienced and the *78%* ($N=7$) are not at all or little experienced.

Table 2 Demographic characteristics

Demographics	Category	%
Gender	Male	67.0%
	Female	33.0%
Grade of teaching	Primary school	11.0%
	Secondary school	78.0%
	Higher school	11.0%
Age	25–29	56.0%
	30–39	22.0%
	50–59	22.0%
Teaching experience (years)	1–5	78.0%
	20–22	22.0%
Teaching specialty	Mathematician	56.0%
	Information technology	22.0%
	Design and Technology	11.0%
	Psychology	11.0%

4.4 Body language score analysis

The overall body language score, and scores for different modalities were recorded for both lectures delivered by each participant. The comparison of the scores between the two lectures allows the derivation of information regarding the impact of the body language analysis application in adjusting the body language of the participants. Figure 5 shows the average quality score per frame for the two participants with the lowest improvement in body language scores, and the two participants with the highest improvement in lecture quality style score. Even for the cases with the lowest improvement, there is noticeable improvement in the quality scores.

Table 3 shows the results of the non-parametric Wilcoxon Sign Rank Test for the educators' group of the differentiation of the scores obtained between the first and the second phase. This test is used because the samples are paired, and the data are in ordinal scale since it has to do with the same persons in different circumstances. The null hypothesis (H_0) is that there is no statistically significant difference for the scores between the first and the second phase. The alternative (H_1) is that there is a statistically significant difference for the scores between the first and the second phase.

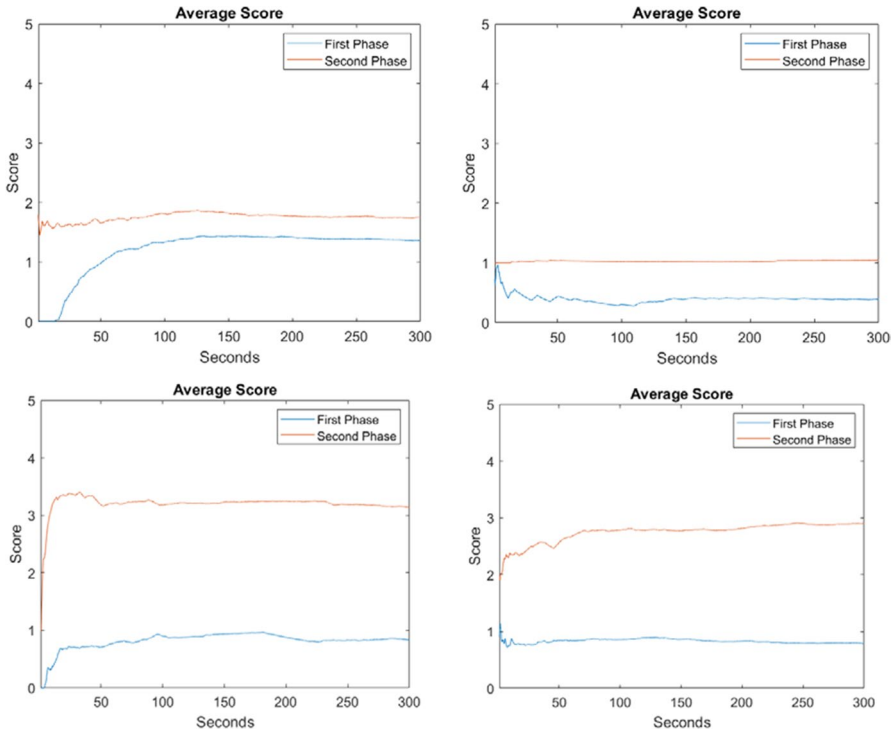

Fig. 5 The average score for two participants with the lowest improvement in body language scores (top row) and two participants with the highest improvement in body language using the BLA application (bottom row)

Table 3 Mean, Standard Deviations, and p-value for the first and second phase regarding the biometric features

Biometric Features	Mean (*Standard Deviation*)	*p*-value
Average Overall Score (first lecture)	0.871 (*0.346*)	**0.011***
Average Overall Score (second lecture)	1.786 (*0.764*)	
Speech Score (first lecture)	0.000 (*0.000*)	0.317
Speech Score (second lecture)	0.125 (*0.354*)	
Hands Score (first lecture)	0.125 (*0.133*)	**0.012***
Hands Score (second lecture)	0.396 (*0.288*)	
Facial Expressions Score(first lecture)	0.165 (*0.093*)	**0.036***
Facial Expressions Score (second lecture)	0.341 (*0.164*)	
Facial Pose Score (first lecture)	0.291 (*0.315*)	**0.025***
Facial Pose Score (second lecture)	0.405 (*0.349*)	
Body Activity Score (first lecture)	0.271 (*0.148*)	0.069
Body Activity Score (second lecture)	0.366 (*0.208*)	

*Statistically significant difference ($p < 0.05$)

Based on the results, there is a statistically significant difference regarding the average score, facial expressions, hand movements, and facial pose. The information provided about the application, and the automated evaluation results presented to the participants after the first lecture, have helped the participants to understand exactly which parameters have been considered 'good' or 'bad' for their teaching. As a result, they have managed to improve the specific metrics of average score, facial expressions, hand movements, body activity and score per frame. Concerning the speech and body activity score, the participants achieved better performance in the second lecture, but there was no statistically significant improvement when compared to the scores of the first lecture.

4.5 Questionnaire results

Table 4 presents the results of reliability analysis for the factor "Satisfactory Feedback" between the first and the second phase, using the Cronbach Alpha coefficient to test internal consistency. Satisfactory values are considered those in the interval [*0.7*,*0.8*) high in the interval [*0.8*, *0.9*) and perfect in the interval [*0.9*,*1.0*] (Nunnaly & Bernstein, 1994). The "Satisfactory Feedback" factor analysis was conducted for the same six questions in both cases (first and second phase). Reliability was perfect for the first phase ($a = 0.946$) and high for the second phase ($a = 0.866$). To test the

Table 4 Reliability analysis for factor "Satisfactory Feedback" before and after intervention

Factor	Items	Cronbach Alpha	Reliability
Satisfactory Feedback (before)	6	0.946	Perfect
Satisfactory Feedback (after)	6	0.866	High

Table 5 Comparisons of application usefulness for improving the quality of the lecture before and after intervention

Question	Mean (*Standard Deviation*)	*p*-value
Do you think the application is useful for improving the lecture quality?	4.78 (0.67)	**0.011***
Do you think the application is useful for improving the lecture quality?	5.67 (0.50)	

*Statistically significant difference ($p < 0.05$)

normality of variables, between the first and the second phase, the Shapiro Wilk test was used, which is considered to have the best results (Razali & Wah, 2011). Normality is accepted only for the factor "Satisfaction from feedback" for the first phase ($p = 0.100$) and second phase ($p = 0.175$).

To test the mean differences between the first and the second phase in related samples (same participants in different conditions), the non-parametric Wilcoxon test was used, in cases that normality was not accepted between the first and the second phase. On the other hand, for the factor "Satisfaction from feedback", the paired samples t-test was used to test mean differences between the first and the second phase because factor is normally distributed in both cases and samples are related. Significance was set at 5% (Field, 2017). According to Table 5, participants agreed more about the application usefulness for improving the quality of the lecture after the intervention than before ($M_{before} = 4.78$ *vs* $M_{after} = 5.67$, $p = 0.011$).

There is a statistically significant difference between the first and the second phase regarding feedback about the facial expressions, body activity, hand movements, facial pose, and average score. According to Table 6, participants presented higher levels of satisfaction from feedback after intervention than before more particular

Table 6 Comparisons of application satisfactory feedback on the teachers' body language before and after intervention

Question	Mean (*Standard Deviation*)	*p*-value
Facial Expressions (before)	4.44 (*1.42*)	**0.014***
Facial Expressions (after)	5.78 (*0.44*)	
Body Activity (before)	4.44 (*1.51*)	**0.041***
Body Activity (after)	5.67 (*0.50*)	
Speech Recognition (before)	4.11 (*1.17*)	0.276
Speech Recognition (after)	4.67 (*1.41*)	
Hand Movements (before)	4.56 (*1.59*)	**0.041***
Hand Movements (after)	5.78 (0.44)	
Facial Pose (before)	4.44(*1.51*)	**0.047***
Facial Pose (after)	5.56(*0.53*)	
Average Score (before)	4.44(*1.42*)	**0.026***
Average Score (after)	5.67(*0.50*)	
Satisfaction from feedback (before)	4.41(1.34)	**0.044***
Satisfaction from feedback (after)	5.52(0.46)	

*Statistically significant difference ($p < 0.05$)

Table 7 Comparisons of application usefulness for improving the quality of the lecture before and after intervention

Question	Mean (*Standard Deviation*)	*p*-value
Do you believe that the feedback provided by the application is useful to improve lecture quality? (Phase 1)	4.10 (0.74)	0.063
Do you believe that the feedback provided by the application is useful to improve lecture quality? (Phase 2)	4.70 (0.48)	

considering facial expressions ($M_{before}=4.44$ *vs* $M_{after}=5.78$, $p=0.014$), body activity ($M_{before}=4.44$ *vs* $M_{after}=5.67$, $p=0.041$), hand movements ($M_{before}=4.56$ *vs* $M_{after}=5.78$, $p=0.041$), facial pose ($M_{before}=4.44$ *vs* $M_{after}=5.56$, $p=0.047$) and average score ($M_{before}=4.44$ *vs* $M_{after}=5.67$, $p=0.026$). All the participants agree that the proposed application provide sufficient feedback regarding the facial expressions, body activity, hand movements, facial pose, and average score. There is no statistically significant difference regarding the speech metric. In the first phase, 33% of educators stated that the system provides much to very much satisfactory feedback, while in the second phase 67% of the participants stated that the system provides much to very much satisfactory feedback.

Even though the mean value related to the participant's impression of the usefulness of the system, increased after the intervention, there is no statistically significant difference between the first and second phase regarding the usefulness of the application in improving lecture quality (see Table 7). Since participants realized that this is a highly useful application even before the intervention, this specific metric received high responses from the first phase of the experiment, the improvement in the score after the intervention was not enough to make the difference statistically significant.

According to Table 8, participants agreed that the application is easy to use in real classroom ($M_{after}=5.44$). Regarding the open-type questions in the questionnaire, the majority of the participants claimed that they preferred the upload option rather than running the application in real time.

Although a small sample was used, the results of the interviews are important as they provide useful feedback regarding the views of the participants about the body language evaluation system.

Table 8 Comparisons of application easy to use before and after intervention

Question	Mean (*Standard Deviation*)
Do you think that the application is easy to use in actual classroom? (Phase 2)	5.44 (*0.73*)

4.6 Semi-structured interview results

In this subsection, the analysis of semi-structured interviews is presented, which was carried out using thematic analysis. Thematic analysis was selected because is considered important to summarize key features. According to Braun and Clarke (2006), thematic analysis is described as "*a method for identifying, analyzing and reporting patterns (themes) within data*".

The present study included the evaluation of the acceptance and impact of the proposed system using questionnaires, interviews as well as the analysis of teachers' lectures. The words, sentences and phrases that had similar conceptual meaning and were important for the research, were classified in groups, while the data of secondary importance have been split from the aforementioned contents. This process is called codification and refers to the process of departmentalization and labelling of texts occurring from interviews (Cresswell, 2002) so as to derive descriptions regarding the impact and acceptance of the system. The responses from the thematic analysis are presented in Table 9. In this study the codification process was performed by a single coder, who repeated the process to confirm the validity and credibility of codification. However, in a future study we aim to perform a more rigorous analysis by involving additional coders to assist in the large data analysis process.

Table 9 Application contribution to the improvement of teachers' body language

Themes	Codes—Quotes
1. App usage opinions	**Satisfaction with its use (1, 2, 3, 4, 5, 6, 7, 8, 9)** *"Yes, very satisfied."* (R1) *"I am satisfied although the results of my lecture were again not as good as I wished."* (R2) *"Yes, it's okay. I really enjoyed the app… I have tried to fix the mistakes I realized I made last time. I did better"* (R3) *"Yes, I am satisfied. Very nice app……. Maybe it could also evaluate the quality of the lesson not just the body language. If only there was an app that could correct our writing even better."* (R4) *"I liked it. I find it a very useful app. "* (R5) *"Yes, I am very satisfied. I definitely have room for improvement…"* (R6) "It is fine. (R7) *"Your idea is very nice…"* (R8) *"I am very satisfied. It is a great application."* (R9) **Improved understanding of results between Phase 1 and Phase 2 (1, 3, 4, 6, 7, 9)** *"Yeah, much better than the first time."* (R1) "*My results are too low. I didn't know what the system evaluates in order to get better results……Yes I have understood them now, but what exactly can I do to get a perfect score?!."* (R3) *"I liked that it brings out the score in every frame. But it is not clear what the number 0 and the number 1 mean… Yes. I tried to improve based on the graphics from last time."* (R4) *"The first time I made the video I didn't understand what 0 was and what 1 stands for."* (R6) "*I can read the results better this time."* (R7) "*Compared to the first phase, I can better understand the graphics. "* (R9) **Understanding results from Phase 1 (2, 5, 8)** *"I have understood them and I will try to work on the points that I had bad results."*(R2) *"I have understood the results. Where my movements were not correct, I got a low score"* (R5) *"As soon as I saw the graphics I understood exactly what the system was measuring because when I saw the asterisks in the video it wasn't clear."* (R8)

Table 9 (continued)

Themes	Codes—Quotes
2. Application contribution to teachers' teaching methods	**Improving facial expressions (2, 3, 4, 5, 6, 7, 8, 9)** *"Yes it was clear at which seconds the facial expressions were not correct."* (R2) *"I did better on the facial expressions compared to the first video."* (R3) *"The app helped me to understand my mistakes…….. I did better on the facial expressions compared to the first video."* (R3) *"I tried to have good expressions this time"* (R4) *"I improved here"* (R5) *"I've improved a lot."* (R6) *"Much better results. Basically, it helped me a lot to be able to improve my expressions."* (R7) *"I find it difficult in a class to have good expressions all the time because there are students who make a fuss……Very good facial expressions."* (R8) *"My facial expressions have improved."* (R9) **No enhancement of facial expressions (1)** *"I can read the graphic but what exactly I did wrong I don't know though*." (R1) **Hand Movement Improvement (2, 4, 5, 6, 7)** *"Yes I understand."* (R2) *"I need to move my arms more."* (R4) *"I did better here too."* (R5) *"Very useful graphics. I didn't know how to move my hands. … Yes I am satisfied with the graphic." (R6)* *"My hand movements are better because now I was moving my hands."* (R7) *"Yes, I understand, I just have to watch the video again to understand even better why in those seconds the hand movements were not so good…"* (R8) *"Quite satisfied."* (R8) *" My performance in hand movements has improved in the second phase. "* (R9) **Improve speech recognition (1, 3)** *"Yes, I understand."* (R1) *"Here I was fine because I managed to make the score 1."* (R3) **Poor speech recognition performance (2, 4, 5, 6, 7, 8, 9)** *"I could probably speak more slowly this time."* (R2) *"Here I didn't understand why I scored 0."* (R4) *"I expected better* " (R5) *"I am not satisfied*" (R6) *"I tried but I had to improve my speaking rate"* (R7) *" It could be better "* (R8) *" I would like it to include more statistics about the voice."* (R9) **Action Improvement (2, 3, 5, 6, 7, 8, 9)** "*Yeah it's clearer now."* (R2) "*I did better*." (R3) *"Much better than last time."* (R5) "*I was good. No complaints.* (R6) "*My actions are more refined."* (R7) "*I was fine compared to the first time*." (R8) " *My performance has improved, but I can do even better.* "(R9) **No Actions improvement (1, 4)** **"***I can read the graphics but what exactly I did I don't know though."* (R1) *"I will try next time to be more focused on the camera"* (R4) **Score improvement for each frame (2, 3, 4, 5, 6, 7, 8, 9)** *"I'm happy compared to the first time…"* (R2) "*And here the score is much better. In some cases I score 4/5."* (R3) *"My score is much better now…"* (R4) *"My score is better now."(*R5) *"Yes, very good"* (R6) *"My score was more improved."* (R7) *"It was fine."* (R8) "*I'm happy compared to the first time."* (R9) **No score improvement for each frame (1)** "*I read the graphic but what exactly I did wrong I don't know though."* (R1)

Table 9 presents the results of participants' views regarding the use of the application. All teachers stated that they were satisfied with the use of the application and the vast majority improved their understanding of the results of the application between Phase 1 and Phase 2.

During the interviews the participants were informed that a value of "0" was defined as "low" body language quality while a value of "1" was defined as "high" body language quality. However, the participants were not provided with information about the way that the actual quality indicators were classified as low and high quality. Therefore, it is reasonable to assume that the improvement in participants' performance is attributed to the overall feedback provided by the application, and not by knowledge regarding the way that specific indicators are classified in low/high quality scores.

Table 10 presents the results regarding the implementation feedback on teachers' body language. The majority of teachers state that there is satisfactory feedback. The rest of the participants mentioned that the feedback provided by the application needs to be presented in a better way.

Table 11 presents the results regarding the usefulness of application in real classroom conditions. Five out of nine participants answered that it needs improvement before the application can be used in real classrooms.

The results of the interviews showed that the majority of teachers are satisfied with the results of the application. In particular, they stated that they had a better understanding of the results of the application in the second phase as opposed to the first phase. Furthermore, the teachers agree that the application improves facial expressions, hand movements, body activity and score per frame. In addition, the majority of teachers state that there is satisfactory feedback with the graphs. Apart from the positive comments, the majority of participants stated that the application needs improvement in order to be used in real classroom conditions and provided highly useful feedback regarding the required modifications.

Table 10 Implementation feedback on teachers' body language

Themes	Codes—Quotes
3. Application feedback	**Satisfactory feedback (1, 2, 6, 7, 8, 9)** *" The application is useful and a good idea. It's a good idea to use graphs for feedback."* (R1) *" Yes, I'm satisfied. "* (R2) *" Yes, I'm very satisfied. I can improve sure. "* (R6) *" I understand the results and I am satisfied. Where my activities were incorrect, I received a low score. "* (R7) *" I liked that it displays the score in every frame. The feedback it provides is useful. "* (R8) *"I'm satisfied with the feedback. "* (R9) **Lack of feedback (3, 4, 5)** *"Yes, I have understood them, but what exactly can I do to get a perfect score?! It would be helpful if you could provide specific feedback."* (R3) *"The app is very useful but it could provide feedback by saying specifically what was wrong without me having to think…… It would be good I think if it also gives feedback which actions are considered good and which are not…… I would like but to provide feedback as to what exactly led me to get a score of 0."* (R4) *"Maybe if there was some feedback to know exactly what to improve."* (R5)

Table 11 Usefulness of application in real classroom conditions

Themes	Codes—Quotes
4. Usefulness of application in real classroom conditions	**Needs improvement in actual classroom conditions (1, 2, 4, 8, 9)** *"The app is useful and a good idea but it needs an explanation of what is good and what helps the lesson. I noticed that my outfit of wearing short pants in class didn't take it into account."* (R1) *"I think if the app recognized the legibility of the teacher's letters it could provide better feedback."* (R2) *"The app is very useful but could provide feedback by specifically stating what was wrong without me having to think."* (R4) *"Quite satisfied……I find it difficult in a class to have good expressions all the time because there are students who make a fuss"* (R8) *"The use of the application is feasible because 0 and 1 make sense…… It would be more useful if colors were added. "* (R9)

5 Discussion

This research examines the differences regarding the body language of educators between two lectures when using an automated body language evaluation application. For this purpose, the experimental procedure consisted of two phases. In the first phase of the experiment, the educators carried out short lectures. During this lecture the body language of the presenters was evaluated based on the automated body language analysis application. Afterwards, participants answered the questions of an online questionnaire that evaluated their experience with the application. Then, a short semi-structured interview followed, a procedure of approximately 15 min with each of the participants in order to discuss the results of the application and explain to them the operation of the application. During the second phase of the experiment participants, they repeated the same procedure as the first phase of the experiment.

The objective of this research is to give answers to three research questions. The research questions and the main conclusions derived are outlined below:

RQ1:*Can this application potentially contribute to the improvement of the educators' body language?*

In the first phase, participants carried out their short lectures without being aware of the metrics that the system took into consideration in order to provide an automated evaluation of their body language. The results regarding interviews, graphics displaying quality metric scores, and questionnaires are presented below.

Interviews: Through the results of the interviews, it was observed that during the first phase the participants were concerned about their performance as they

anticipated better results because they did not know which metrics are considered "good" by the system in order to acquire a good evaluation. On the other hand, during the second phase of the interviews the participants were satisfied with the results once they understood the function of the application that encourages them to adopt improved body-language, and for this reason an upgrade on the performance indicators was observed.

Graphics displaying quality metric scores: The analysis of the graphical results (see Fig. 5) showed that the application contributed to the improvement of the educators' teaching style body language regarding the average score, facial expressions, facial pose and hand movements. Concerning the speech and body activity score, the participants achieved better performance in the second lecture, but there was no statistically significant improvement when compared with the results of the first lecture.

Questionnaire: Despite the fact that no statistically significant difference was recorded for the "body activity" metric, according to the responses registered after the intervention, the participants considered the "body activity" metric important, in line with the relevant literature (Tok & Temel, 2014), hence it is crucial to include the "body activity" metric in the metrics under evaluation. For the speech metric there was no statistically significant difference because the participants presented difficulty to comply with the standards that were defined for word density, speed and intonation. The non-statistically significant difference before and after the intervention concerning the speech metric is due to the fact that the educators in this study were not accustomed to speaking at a normal pace and not monotonously in their lectures, and as a result participants encountered difficulty in improving the speech metric. According to Najmiddinova (2024), a monotonous voice combined with variations in speed results in students not maintaining their attention. In future research, we aim to train educators to deal in a better way with the metrics where they did not show significant improvement, such as the speech metric. Furthermore, the results of the questionnaire showed that the participants considered the application to be useful for improving their teaching style body language.

Body language plays a crucial role in communication, as it can convey emotions, engagement, and enthusiasm, which are essential for effective teaching (Tok & Temel, 2014). Furthermore, assessing body activity provides insights into the level of teacher involvement and interaction with students during instruction, influencing student engagement and comprehension hence body language can serve as a non-verbal cue for classroom management and establishing a positive learning environment. Additionally, the validity and reliability of measuring body activity contribute to understanding its impact on teaching effectiveness, emphasizing the importance of this metric in evaluating and improving instructional practices. Overall, monitoring body activity in teaching facilitates better understanding and enhancement of teacher-student communication, engagement, and overall teaching quality. Therefore, even though in the experiments no significant variation in the body activity score was observed, it is imperative that this metric is used in the estimation of the overall body language quality score. Therefore, using this system, educators could be trained to use their own body language properly and effectively during lectures. According to Gulec and Temel (2015), educators must use their body language in

an effective way. Doing so, they will be able to provide a more charming and sufficient learning environment, in addition to enhancing the communication and the understanding of students (Tai, 2014) as well as affecting their performance and evaluation.

Based on the results of the experimental evaluation, as documented in the previous paragraphs, the first research question has a positive answer. The application can potentially contribute to the improvement of teachers' teaching style body language, that could contribute to lecture quality improvement.

RQ2:*Does the application provide sufficient feedback regarding the body language of educators?*

According to the results of the interviews and the open type questions of the questionnaire, the participants claimed that the application is useful, helpful and that it provides sufficient feedback. The questionnaires results confirm the interviews' results because there was a statistically significant difference between the first and second phases. The participants agree that the application can provide satisfactory feedback regarding the facial expressions, body activity, hand movement, facial pose, average score. Furthermore, the participants agree that the feedback provided by the application is useful for improving the quality of the lecture. In order for the system to provide better feedback, it was suggested that the application should contain further explanations that defines Right and Wrong body language practises. Furthermore, the participants mentioned that it would be preferable for the users to have the choice of which metric (or metrics) the system considers. Based on the results of the experimental evaluation the second research hypothesis was confirmed. The application provides satisfactory feedback on the teachers' body language, considering facial expressions, body activity, hand movements, facial pose, facial expressions, score per frame and average score.

RQ3:*Is the use of this application acceptable in actual classroom conditions?*

In the interviews, all educators claimed that they understood the functions of the application. At the end of the interviews participants were asked if they would prefer to use the application in real time or to upload a video after the end of the class and get the results after the lecture. The majority of the participants preferred the upload option rather than running the application in real time because they believed that real time operation might affect their focus during class and the interaction between educators and students. The majority of the participants stated that the application needs improvement before applying it in real classrooms. They mentioned that it would be better if there was further specific information regarding their performance without having to understand the graphics. Moreover, they would like the application to recognise their writing and their outfits and provide additional statistics regarding speech recognition. Furthermore, in real classroom conditions they mentioned that the application should be able to inform the educators with a screen alert about their performance quality (Right / Wrong) in order to avoid the reading of graphics during class and thus keep the student—educator interaction undisturbed. In real classroom

conditions, the lecture could be recorded and after the end of the class the educator would receive the feedback and not in real time because their focus on teaching would be affected. Regarding the third research hypothesis, the proposed application can be used at the end of the lesson by teachers for self-assessment regarding their body language and their speaking skills, but at the same time of number of desirable improvements were defined.

Although the results of the experiments proved the potential of the application in improving lecture delivery practises, further experimentation can enhance the significance of the results by dealing with the limitations of the current study. For example, in the future it is desirable to assess the impact of the application with more educators and assess the application on a full series of lectures delivered over a whole semester. However, time constraints related to the completion of the project in combination with long procedures to get permissions to run the experiments in real classes for a whole semester, did not allow the implementation of such large-scale experiments for the current study.

The present study aims to fill the gap in the literature regarding the acceptance of body language assessment systems by stakeholders. The results of the study show that educators had a positive response to the proposed system. This indicates that educators understand the impact of body language quality on the learning process. This is in line with recent studies that delve into the influence of body language on the learning process. Khuman (2024) conduct an in-depth literature review to explore the impact of non-verbal communication in the teaching learning. This study delves into foundational theories, explores the various implications, challenges in the effective use of nonverbal communication, and suggests recommendations for professional development of educators. Furthermore, Najmiddinova (2024) investigate the dynamics of the teacher's body language and voice in the learning process. The authors state that teachers who use effective body language can create a positive environment and communicate better with their students. Similarly, Toshpo'latova (2024) highlights the importance of gestures and body language in English teaching. This study show that non-verbal communication improves students' understanding.

6 Limitations

Although the results of the experiments proved the potential of the application in improving lecture delivery practices, further experimentation can enhance the significance of the results by dealing with the limitations of the current study. For example, in the future, it is desirable to assess the impact of the application with more educators and assess the application on a full series of lectures delivered over a whole semester. However, time constraints related to the completion of the project in combination with long procedures to get permission to run the experiments in real classes for a whole semester, did not allow the implementation of such large-scale experiments for the current study.

Another limitation of the present study concerns the fact that the criteria for low/high quality body language are permanently fixed (see Table 1). However, depending on the nature of different classes, and the conditions prevailing during class delivery

it is possible that the optimum body language indicators can be differentiated. For example, although a 'sad' expression is generally considered as a low-quality body language feature, in cases that a lecturer needs to convey messages of disappointment towards the audience, a 'sad' expression is more appropriate for a lecturer. Ideally the set of quality indicators should be adaptive based on other factors related to the audience, class settings and the content of the lecture. Given the acceptance of lecturers in the pilot system evaluated in this work, as part of our future work in the area we plan to deal with the issue of adaptive and continuous adjustment of quality indicators throughout class delivery. This line of work, requires the provision of information related to the class content, and the use of additional sensors for monitoring the class environment. It should be noted though that the extra benefits of using adaptive quality indicators will come at the cost of decreased versatility and ease of use of the body language assessment system.

7 Conclusions and future works

The current paper presents an experimental investigation regarding the impact and acceptance of an automated Body Language Analysis (BLA) application. The BLA application is an expert system that imitates human expertise for assessing the quality of the teachers' body language. The BLA application requires only a personal computer equipped with a camera and microphone to run and hence it does not require any specialized equipment allowing educators to use the application during lectures supporting in that way continuous personal self-evaluation and improvement. The results of the study showed the potential of the proposed system in supporting educators to improve their body language during lecture delivery. Furthermore, based on feedback received by educators, the application provides satisfactory feedback on the teachers' body language, considering facial expressions, body activity, hand movements, facial pose, facial expressions, and the overall quality score. Moreover, the results showed that educators consider the proposed application easy to use as it does not require any specialized equipment. Additionally, the participants agree that the proposed application can be used at the end of the lesson by teachers for self-reflection and assessment. Teachers believe that the application is very useful as it can help them improve their body language during lecture delivery, thus enhancing the overall quality of their lectures by providing an enhanced learning experience for students.

In the future, we aim to introduce the BLA application in real classrooms and assess its performance over a series of lectures delivered by a large number of educators. In parallel, feedback received during the evaluation will guide further improvements to the systems. For example, we plan to improve the interface and feed visualization so that users get feedback in a user-friendly manner. Also, in the future we will consider the use of additional cameras and/or wearable cameras pointing at the educator, to cover the movements of educator in class rather than using a single static camera that captures images only at a certain location. Further study is also required to make sure that the use of the BLA application in real school environments does not violate any data privacy rules of educators and students.

Furthermore, in the future we aim to perform experiments to redefine the scores corresponding to low/high quality speech as those in the literature may not be accurate. In addition, we plan to provide training to teachers in the metrics that present particular difficulty such as for example for the speech metric. In future research, we aim to perform analyses for each teaching specialty and conduct comparative assessments between groups so that the applicability of the proposed system for the delivery of different subjects, and different levels is determined. In the future, we aim to examine the use, impact, and operation of the proposed system in other potential areas besides education. For example, the proposed teaching style evaluation methodology can be used for assessing the body language of sales personnel. By analysing non-verbal cues like gestures and tone of voice, the system can identify effective sales techniques and areas for improvement, enhancing communication and rapport-building skills. The commercial exploitation of the system will also be going to be addressed in the future.

Acknowledgements This project has received partial funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 and the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination and Development.

Funding Open access funding provided by the Cyprus Libraries Consortium (CLC).

Data availability The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Consent to publish I understand that the text and any pictures published in the article will be freely available on the journal and may be seen by the general public.

Competing interests The authors have no financial or proprietary interests in any material discussed in this article.

References

Abdulrahaman, M. D., Faruk, N., Oloyede, A. A., Surajudeen-Bakinde, N. T., Olawoyin, L. A., Mejabi, O. V., ... & Azeez, A. L. (2020). Multimedia tools in the teaching and learning processes: A systematic review. *Heliyon*, *6*(11). https://doi.org/10.1016/j.heliyon.2020.e05312

Albergaria-Almeida, P. (2010). Classroom questioning: Teachers' perceptions and practices. *Procedia-Social and Behavioral Sciences, 2*(2), 305–309.

Archer, J., Cantrell, S., Holtzman, S. L., Joe, J. N., Tocci, C. M., & Wood, J. (2016). *Better feedback for better teaching: A practical guide to improving classroom observations*. John Wiley & Sons.

Azer, S. A. (2005). The qualities of a good teacher: How can they be acquired and sustained? *Journal of the Royal Society of Medicine,98*(2), 67–69.

Bambaeeroo, F., & Shokrpour, N. (2017). The impact of the teachers' non-verbal communication on success in teaching. *Journal of Advances in Medical Education & Professionalism, 5*(2), 51.

Barmaki, R., & Hughes, C. (2018). Gesturing and embodiment in teaching: Investigating the nonverbal behavior of teachers in a virtual rehearsal environment. In *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1).

Behar, L. S., & George, P. S. (2013). Teachers' use of curriculum knowledge. In A. C. Ornstein, & L. S. Behar (Eds.), *Our evolving curriculum* (pp. 48–69). Routledge.

Benzer, A. (2012). Teachers' Opinions about the Use of Body Language. *Education, 132*(3), 467–473.

Bhatia, M., & Kaur, A. (2021). Quantum computing inspired framework of student performance assessment in smart classroom. *Transactions on Emerging Telecommunications Technologies,32*(9), e4094.

Bower, M. G., Moloney, R. A., Cavanagh, M. S., & Sweller, N. (2013). Assessing Preservice Teachers' Presentation Capabilities: Contrasting the Modes of Communication with the Constructed Impression. *Australian Journal of Teacher Education,38*(8), 111–130.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology,3*(2), 77–101.

Caswell, C., & Neill, S. (2003). *Body language for competent teachers*. Routledge.

Corley, M. A., & Rauscher, C. (2013). *12: Deeper Learning through Questioning*. Teaching Excellence in Adult Literacy.

Cresswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative*. Prentice Hall.

Darling-Hammond, L., Newton, X., & Wei, R. C. (2010). Evaluating teacher education outcomes: A study of the Stanford Teacher Education Programme. *Journal of education for teaching, 36*(4), 369–388.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (2013). Teacher evaluation in the organizational context: A review of the literature. *New directions in educational evaluation* (pp. 203–253).

Denham, M. A., & Onwuegbuzie, A. J. (2013). Beyond words: Using nonverbal communication data in research to enhance thick description and interpretation. *International Journal of Qualitative Methods,12*(1), 670–696.

Dimitriadou, E., & Lanitis, A. (2023). An integrated framework for developing and evaluating an automated lecture style assessment system. *arXiv preprint arXiv:2312.00201*.

Dolan, R. (2017). Effective presentation skills. *FEMS Microbiology Letters,364*(24), fnx235.

Dubinsky, J. M., Roehrig, G., & Varma, S. (2022). A place for neuroscience in teacher knowledge and education. *Mind, Brain, and Education,16*(4), 267–276.

Fast, J. (1970). *Body Language*. M. Evans & Co.

Field, A. (2017). *Discovering statistics using IBM SPSS* (5th ed.). Sage Publications Ltd.

Gelula, M. H. (1997). Effective lecture presentation skills. *Surgical Neurology,47*(2), 201–204.

Gulec, S., & Temel, H. (2015). Body language using skills of teacher candidates from departments of mathematics education and social studies education. *Procedia-Social and Behavioral Sciences,186*, 161–168.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research,77*(1), 81–112.

Hunter, E. J., & Titze, I. R. (2010). Variations in intensity, fundamental frequency, and voicing for teachers in occupational versus nonoccupational settings. https://doi.org/10.1044/1092-4388(2009/09-0040)

Hussain, A., Badshah Rehman, D. M. M. S., Abdul Naseer, D. S. B., & Muhammad, A. (2022). Introduction and development of body language and its importance in education (An Overview In The Context Of Contemporary And Islamic Teachings). *Journal of Positive School Psychology,6*(10), 4355–4362.

Jensen, E., Dale, M., Donnelly, P. J., Stone, C., Kelly, S., Godley, A., & D'Mello, S. K. (2020). Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).

Jensen, E., L. Pugh, S., & K. D'Mello, S. (2021). A deep transfer learning approach to modeling teacher discourse in the classroom. In *LAK21: 11th international learning analytics and knowledge conference* (pp. 302–312).

Karaca, Y., & Filiz, B. (2023). Examination of the body language competencies of physical education teachers. *The Physical Educator, 80*(2), 212–235.

Khuman, P. (2024). The impact of non-verbal communication in teaching: Enhancing educational effectiveness.

Kucuk, T. (2023). The power of body language in education: A study of teachers' perceptions. *International Journal of Social Sciences and Educational Studies, 10*(3), 275–289.

Lasri, I., Solh, A. R., & El Belkacemi, M. (2019). Facial emotion recognition of students using convolutional neural network. In *2019 third international conference on intelligent computing in data sciences (ICDS)* (pp. 1–6). https://doi.org/10.1109/ICDS47004.2019.8942386

Liakopoulou, M. (2011). Teachers' pedagogical competence as a prerequisite for entering the profession. *European Journal of Education,46*(4), 474–488.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature,244*, 522–523.

Miller, P. W. (2005). Body language in the classroom. *Techniques: Connecting Education and Careers,80*(8), 28–30.

Mohammadreza, E., & Safabakhsh, R. (2021). Lecture quality assessment based on the audience reactions using machine learning and neural networks. *Computers and Education: Artificial Intelligence,2*, 100022.

Najmiddinova, S. N. (2024). The comprehensive influence of body language and teacher voice in classroom communication. In *International scientific conference " innovative trends in science, practice and education"* (Vol. 3, No. 2, pp. 27–32).

Nunnaly, J., & Bernstein, I. (1994). *Psycometric theory* (3rd ed.). McGraw-Hill Inc.

Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies,27*(6), 7893–7925.

Pease, A. (1981). *Body language: How to read others' thoughts by their gestures*. Sheldon Press.

Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign language recognition: A deep survey. *Expert Systems with Applications,164*, 113794.

Razali, N., & Wah, Y. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics,2*, 21–33.

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly,17*(2), 219–240.

Rosen, R. S., Turtletaub, M., DeLouise, M., & Drake, S. (2015). Teacher-as-researcher paradigm for sign language teachers: Toward evidence-based pedagogies for improved learner outcomes. *Sign Language Studies,16*(1), 86–116.

Saroyan, A., & Snell, L. S. (1997). Variations in Lecturing Styles. *Higher Education, 33*(1), 85–104.

Schempp, P. G., Manross, D., Tan, S. K., & Fincher, M. D. (1998). Subject expertise and teachers' knowledge. *Journal of Teaching in Physical Education, 17*(3), 342–356.

Schussler, D. L., Stooksberry, L. M., & Bercaw, L. A. (2010). Understanding teacher candidate dispositions: Reflecting to build self-awareness. *Journal of Teacher Education,61*(4), 350–363.

Short, F., & Martin, J. (2011). Presentation vs. Performance: Effects of lecturing style in higher education on student preference and student learning. *Psychology Teaching Review,17*(2), 71–82.

Tai, Y. (2014). The application of body language in English teaching. *Journal of Language Teaching and Research,5*(5), 1205.

Tok, M., & Temel, H. (2014). *Body language scale: Validity and reliablity study*. Eğitimde Kuram ve Uygulama.

Toshpo'latova, Z. (2024). Using gestures and body language in efl classes. In *Conference Proceedings: Fostering Your Research Spirit* (pp. 239–241).

Turaev, S., Al-Dabet, S., Babu, A., Rustamov, Z., Rustamov, J., Zaki, N., ... & Loo, C. K. (2023). Review and analysis of patients' body language from an artificial intelligence perspective. *IEEE Access*.

Van Klaveren, C. (2011). Lecturing style teaching and student performance. *Economics of Education Review,30*(4), 729–739.

Wang, L. (2021). British English-speaking speed 2020. *Academia Journal of Humanities & Social Sciences,4*, 93–100.

Winarno, W. (2017). Design and implementation of web-based lecture evaluation system. *Journal Pendidikan Islam UIN Sunan Gunung Djati,3*(2), 235–248.

Woolfolk, A. E., & Brooks, D. M. (1985). The influence of teachers' nonverbal behaviors on students' perceptions. *Journal of Educational Psychology,77*(4), 425–435.

Yang, W. J., Zhou, Y. J., & Yuan, S. (2012). Study of teaching assessment based on BP neural network. *Advanced Materials Research,524–527*, 3861–3865. https://doi.org/10.4028/www.scientific.net/AMR.524-527.3861

Zhao, Y., & Tang, W. (2019). Modeling and analysis of college teaching quality based on Bp neural network. In *3rd International conference on advancement of the theory and practices in education (ICATPE 2019.).*

Zhu, H. (2022). English teaching quality evaluation based on analytic hierarchy process and fuzzy decision tree algorithm. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2022/5398085

Authors and Affiliations

Eleni Dimitriadou[1,2] · **Andreas Lanitis**[1,2]

✉ Eleni Dimitriadou
ela.dimitriadou@edu.cut.ac.cy

Andreas Lanitis
andreas.lanitis@cut.ac.cy

[1] Visual Media Computing Lab, Department of Multimedia and Graphic Arts, Cyprus University of Technology, Limassol, Cyprus

[2] CYENS Centre of Excellence, Nicosia, Cyprus