

Opportunities and Challenges of AI to Support Student Assessment in Computing Education: A Systematic Literature Review

Simone C. dos Santos^a and Gilberto A. S. Junior
Centro de Informática, Federal University of Pernambuco, Recife, Brazil

Keywords: Computing Education, Student Assessment, Artificial Intelligence, Systematic Literature Review.

Abstract: This study investigates how Artificial Intelligence (AI) can support student assessment in computing education through a systematic literature review of twenty studies from the past decade. AI's evolution has significantly impacted various fields, including education, offering advanced capabilities for personalized teaching, continuous evaluation, and performance prediction. Analysing these studies, evidence showed a focus on undergraduate students and the employment primarily face-to-face teaching methods, with engineering education and serious games being more cited contexts. These studies also reveal AI's potential to create personalized learning experiences using techniques like fuzzy logic, KNN algorithms, and predictive models to analyse student interactions and performance, particularly in educational games and online courses. The positive findings demonstrate AI's effectiveness in classifying students' learning profiles, predicting employability, providing real-time assessments, facilitating targeted interventions, and improving learning outcomes through personalization. Automated assessments via AI have been shown to reduce teachers' workload by offering accurate, real-time feedback. However, the studies also highlighted challenges concerning student engagement, teacher material quality, model generalization, and technical obstacles such as natural language processing, algorithm stability, and data cleaning. These data-driven factors emphasize the necessity for further advancements in AI to enhance continuous and effective student assessment as part of the personalized learning process.

1 INTRODUCTION


Computing teaching has faced many challenges, especially in promoting more effective learning methods and enabling students to solve practical issues (Computer Science Curricula, 2020). According to Hoed (2016), the high theoretical content of disciplines is not very stimulating and leads students to try memorizing specific contents, but not always successfully. These aspects contribute to many students dropping out in the initial periods of undergraduate courses.

To anticipate learning difficulties, intelligent tools can help teaching methods by providing, for example, personalized teaching according to the needs of each student (Hull and Du Boulay, 2015). In addition to assisting in the learning process, these tools can support teachers during assessments and learning management (Zafar and Albidewi, 2015; Malik et al., 2022), monitor student performance, and carry out

appropriate interventions throughout the teaching process (Maia & Santos, 2022).

According to Salem (2011), the use of AI in education can be divided into seven main research areas: educational systems, teaching aspects, learning aspects, cognitive science, knowledge structure, intelligent tools and shells, and educational interfaces. Each research area has different applications, such as intelligent tutoring systems, educational robots, and assessment systems. Specifically for teaching computing, several studies and research have used AI to make the learning process more efficient (Broisin et al., 2017; Mostafavi and Barnes, 2017; Rvers and Koedinger, 2017).

Among several possibilities of use, an AI can process much more data than a human being and can provide more personalized learning through intelligent tutors that can, for example, detect whether a student learns better with failures or tips (Salem, 2011). Therefore, this technology has enormous

^a <https://orcid.org/0000-0002-7903-9981>

potential to support student assessment during the learning process and promote improvement interventions before a course or discipline ends when it may already be too late (Cope et al., 2021). In addition, AI can be used in learning management to predict the student's final performance and, using their history, indicate subjects that best fit their needs (Salem, 2011).

Because it is an area of vast knowledge and possibilities for diverse applications, choosing the most appropriate technologies for the desired objective is essential, as highlighted in (Mazza and Milani, 2005). At this point, understanding the potential of AI to name, calculate, measure, and represent data, giving meaning to its use (Cope et al., 2021), can bring more clarity about the opportunities and challenges to be overcome.

In this context, the following central research question motivated this study: RQ) *How can AI support student assessment in computing education?* To answer this question, this study used Kitchenham's Systematic Literature Review (SLR) method (Kitchenham and Charters, 2007), selecting primary studies from 2012 to 2022 on the research basis of great relevance for Computing Education Research.

To report the research results, this paper is divided into six sections. After this brief introduction, Section 2 presents the primary backgrounds for the research and its analysis. Section 3 describes the SLR method. Section 4 presents the results, discussed in Section 5. Finally, Section 6 comments on the conclusions and future works.

2 BACKGROUND

2.1 Computing Education Challenges

Considering education as a process that transforms those who learn (Creasy, 2018), students' engagement in this process is essential. However, student engagement is regarded as one of the biggest challenges in higher education (Quaye et al., 2019), which is no different in computing education. According to Hoed (2016), the difficulty in abstracting the content of the subjects in the initial periods of computing courses is a critical factor in discouraging students from continuing the course. For example, subjects focused on teaching calculus and algorithms often represent bottlenecks for first-year students.

The difficulty in assimilating the content is often linked to the student's habit of memorizing the subject in the classroom, motivated by a content-based

assessment process focused on punishing rather than stimulating learning. As a result, failure in initial of disciplines is one of the biggest causes of dropout in computing courses.

In this scenario, new teaching methodologies and practices, aligned with assessment models and continuous feedback, can be allies to overcome many challenges. According to Vihavainen et al. (2011), assessment based on constant feedback helps in computing learning, providing students with greater motivation and self-efficacy and, consequently, more significant engagement. As it requires a lot of effort, this type of assessment is usually supported by systems and information technology, especially in the case of highly populated classes.

On the one hand, Yadav (2016) discusses the difficulties of teaching computer science in a school context, highlighting student assessment as one of the main challenges. Among the points mentioned, the difficulty in accurately measuring students' learning is worth highlighting due to the lack of tools for evaluating computational exercises. Furthermore, it is noted that, given the interdisciplinary nature of the questions, the accurate evaluation of these exercises requires a lot of effort. As a result, teachers use models for assessment purposes, which can reduce students' creativity in carrying out tasks. On the other hand, Hull and Du Boulay (2015) used an Intelligent Tutor System (ITS) to provide continuous and personalized feedback according to students' current knowledge in an SQL course. As a result, they found that students who used the STI had better results than those who did not. These studies reinforce the importance of the assessment modalities, considering different aspects and models.

2.2 Assessment Modalities

It is possible to implement assessments in three primary modalities: *diagnostic*, *formative*, and *summative*.

Diagnostic assessment aims to identify students' prior knowledge, skills, strengths, and areas for improvement before instruction begins (Huhta, 2008). This information helps educators tailor their teaching strategies to meet the specific needs of their students. As main characteristics, we can point out that it is usually conducted at the beginning of a course or unit; it is not typically graded; instead, it serves as a tool for instructional planning, helping set realistic learning goals and benchmarks for improvement (Huff & Goodman, 2007). Examples of diagnostic assessment are pre-tests that assess students' knowledge of a subject before starting a new

unit or course; surveys or questionnaires that gather information about students' learning preferences and study habits; skill assessments that evaluate students' competencies in specific areas, such as writing or math, to tailor instruction accordingly (Gorin, 2007).

Formative assessment is designed to provide continuous feedback and information during the instructional process rather than at the end (Huhta, 2008). The goal is to monitor student learning and provide ongoing feedback that instructors can use to improve their teaching and students to improve their knowledge (Gallardo, 2021). This type of assessment is usually informal and can be integrated into daily teaching activities. It helps identify students' strengths and weaknesses in real time, allowing for immediate adjustments in teaching methods (Kemp & Scaife, 2012). It also encourages student involvement in their learning process through self-assessment and peer feedback. Some examples are quizzes and short tests that are not graded or have a low impact on the final grade; classroom discussions where students are encouraged to ask questions and express their understanding; homework assignments that provide insights into students' progress; peer review sessions where students critique each other's work (Bennett, 2011).

Summative assessment evaluates student learning, knowledge, proficiency, or success after an instructional period (Huhta, 2008). This type of assessment is typically used to assign grades and measure achievement against predefined standards. This assessment is usually formal and structured, often high stakes, affecting students' grades or progression. It also provides a way to compare student performance across different educational settings. More familiar examples of this kind of assessment are final exams that cover the content taught throughout the course, term papers or research projects that require students to demonstrate their understanding and synthesis of the material, and standardized tests that measure student performance against national or state benchmarks (Gallardo, 2021).

Each assessment type plays a crucial role in the educational process, providing valuable information that can help improve teaching strategies and enhance student learning outcomes (Wiliam, 2000).

2.3 AI Technology in Education

According to McCarthy (2010), Artificial Intelligence (AI) "is a science in which intelligent programs are produced that can be used to understand human intelligence." At the same time, artificial

intelligence is inferior to human intelligence, as it only performs calculations; still, it is superior, considering these calculations are made at high speed and with large numbers (Cope et al., 2021). To make sense of AI calculations, Cope et al. (2021) point out four transpositions between number and meaning to determine what is possible to achieve with the application of Artificial Intelligence (AI) technology in education: Naming, Calculability, Measurability, and Representability. AI for Naming is usually related to the ontology of classifying data (Novak, 2010). AI to Calculate is determined by the algorithm used, identifying what can be calculated, whether it is possible to use a large amount of data, or whether human intervention is required (Gulson and Sellar, 2022). Measurability defines how data will be collected, for example, through short answers, questionnaires, and essays, or still with automated technology such as data sensors, motion detectors, keystroke counters, clickstream records, engagement log files, virtual and augmented reality pathways, or QR code scans (Montebello, 2019). Finally, the transposition of representativeness is defined as the evidence of meaning materialized in text, image, and sound, which is how the result of a model is represented to users (Zhang et al., 2020).

Most definitions of artificial intelligence are focused on its methods, assuming that the evolution of these calculation methods can make AI results more human-like. The best-known methods are machine learning, deep learning in neural networks, and quantum computing. According to (Cope et al., 2021), these methods cannot exceed the limits of the four mentioned transpositions, considering that AI involves more than these methods.

Machine Learning uses statistical methods to make predictions based on observed patterns. When an image or text is tagged or classified using labels applied by human "trainers," this method is said to be Supervised. In unsupervised machine learning, the computer identifies statistical patterns, and human trainers are asked to label the text or images where these patterns occur (Zhai & Massung, 2016: 34–6).

Deep Learning and Neural Networks are multilayer statistical sequences identifying patterns in patterns (Krizhevsky, Sutskever, & Hinton, 2012; Rumelhart, Hinton, & Williams, 1986). To function, they require large amounts of data and computational processing. Multiple layers of network analysis produce less intuitively explainable results than the single-layer patterns of first-order machine learning.

Quantum computing still holds promise, applying ideas from quantum mechanics to computation and replacing bits of 0 and 1 with qubits, determinable as

probabilities rather than defined numbers (Feynman, 1982; Harrow, 2015).

It is important to emphasize that Artificial Intelligence will never be the same as human intelligence. AI cannot replace teachers but support them through new educational models, using its full potential to help them (Cope et al., 2021).

2.4 Related Works

The authors of the current study are part of a research group focused on innovative experiences in computing education called NEXT (innovative educational experiences in technology). To investigate the use of innovative strategies aimed specifically at student assessment, the group initially carried out a systematic review with a focus on student assessment models in the active learning contexts and, after that, the group carried out an ad hoc investigation for works that involved AI (due to its growing impact on society) and student assessment, motivated for the interest of a group member concluding his undergraduate course in computing engineering. From this initial research step, four related works were analysed.

In (Lopes & Santos, 2021), the authors presented a Systematic Literature Review on student assessment models in the context of problem-based learning (PBL). Of the 47 studies selected, the authors identified that most studies use a conventional assessment model, focusing on learning technical content. In PBL, the authors observed that, in addition to better absorbing technical knowledge, students could develop soft skills, explore creativity, optimize interpersonal relationships in group work, explore the active resolution of practical problems, and leave the conventional learning environment. However, the study found few assessment models prepared for this range of skills and provided no evidence of using AI to support these models.

The study in González-Calatayud et al. (2021) analyses the use of AI for student assessment using the RSL method in the Scopus and Web of Science databases. Of the 454 articles initially found, 22 studies are selected with a focus on identifying the educational and technological impacts of the use of AI and the type of assessment enhanced by AI in any educational context. The authors conclude that AI has many possibilities, mainly in tutoring, assessment, and personalization of education. Still, they also point out several challenges, such as the need to humanize technology, which involves the preparation of IT professionals, students, and teachers for future education transformation. The current study focuses

on identifying the technological applicability of AI to support student assessment models in computing education. Therefore, it addresses a different perspective from the study in (González-Calatayud et al., 2021), which focuses on the educational impact of the application of AI.

The study (Loras et al., 2021) presents a Systematic Literature Review to determine what is known about the study behaviours of computing students and the role that educational projects play in their training. After applying all criteria, 107 studies were selected. The authors identified a common tendency for students to focus only on a specific subject, with introductory programming courses predominating. This study did not consider the use of technology to support student assessment.

The systematic mapping study in (Ouyang et al., 2022) examines the functions of AI, technologies used, and overall effects in 32 articles published between 2011 and 2020 focusing on online higher education. The results show the functions of AI, emphasizing predicting the state of learning, performance or satisfaction, resource recommendation, automatic assessment, and improvement of the learning experience. The study also identifies that traditional AI technologies are usually adopted (decision trees, neural networks, machine learning), while more advanced techniques (e.g., genetic algorithm, deep learning) are rarely used. The effects generated by AI applications highlight the quality of prediction and recommendations and improvement in students' academic performance and their online engagement.

3 RESEARCH METHOD

According to Kitchenham and Charters (2007), a Systematic Literature Review (SLR) is a method designed to identify, evaluate, and interpret all available research relevant to a particular research question, area, topic, or phenomenon of interest. This study used the SLR method in three phases to conduct this SRL: Planning, Conducting, and Reporting the review.

In the first phase, some activities were realized to have a better understanding of the problem, such as its context (computing education challenges), motivation (AI technology in education), and research concerning student assessments (related work), as discussed in Section 2. Returning to the central research question of the current study "*How can AI support student assessment in computing education?*", we defined two secondary questions to

guide the search and selection processes of primary studies:

- RQ1) What are the **opportunities** for using AI in assessing students in computing, considering these contexts (type of use and impact)?
- RQ2) What are the main **challenges** of using AI from the selected contexts?

To answer these research questions, a generic search string was used, including search terms concerning AI technologies, student assessment, and practical characteristics of the studies, as shown in Table 1.

Table 1: Search string.

Generic String
("AI" OR "Artificial intelligence" OR "machine learning" OR "data mining") AND ("student assessment" OR "student evaluation") AND ("experience" OR "case study" OR "experiment")

The search string was tested using an iterative approach and refined with the help of two researchers: one graduating student in computing engineering, a researcher in AI, and one DSc., a researcher in computing education.

With this string, the collection process was carried out by extracting primary studies from four highly relevant research bases with a high impact factor and a wide variety of studies in computing education: IEEE Transaction on Education (ToE), ACM Transactions on Computing Education (ToCE), Wiley Computer Applications in Engineering Education (CAEE), and Education Resources Information Centre (ERIC).

The search process was only automated through the respective base engine, restricted to the range of papers published from 2012 to 2022 (closed to new publications). It is important to emphasize that, despite the study focusing on computing education, this term can vary enormously in the experience reports on the subject, indicating, for example, disciplines, specific courses, and environments. For this reason, we chose not to include terms related to this domain, leaving this aspect as one of the selection criteria.

To conduct this RSL, a three-step procedure was carried out to select the articles that could answer the research questions. First, articles from 2012 to 2022 were selected according to the research planning, resulting in 90 articles. At this stage, 20 studies were discarded for being outside this period, resulting 70 studies. Second, by reading the titles and abstracts, most of these articles were discarded based on exclusion criteria, resulting 30 studies. We defined

the following exclusion criteria: i) Lack of alignment with the research theme; ii) Secondary studies (another RSL or MS); iii) Articles with paid content; iv) Duplicate or similar articles; v) Articles unavailable for download or viewing; vi) Articles with less than four pages; vii) Articles present in book; viii) Articles not included in conference proceedings. It is important to note that, due to the nature of the research databases, all articles are written in English. At the end of the selection procedure, the following quality criteria were applied: i) Well-defined methodology; ii) Proposal well presented; iii) Practical application; iv) Commented research limitations and threats; v) Completeness and Clarity of content. Each quality criterion received an evaluation, following the value scale: 0 (no attend), 0.5 (partially), and 1 (attend). The studies with a score lower than 3 were excluded from the research, resulting 20 qualified primary studies, as shown in Appendix A. The PRISMA flow chart of study selection process is shown in Fig. 1.

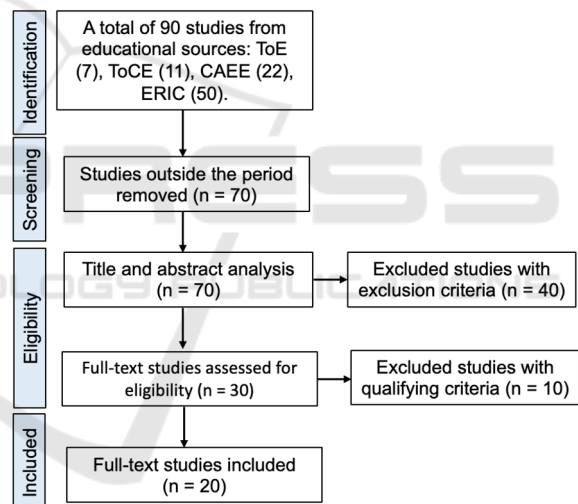


Figure 1: PRISMA flow chart of study selection process.

Fig. 2 shows results considering each research source at the end of the selection process.

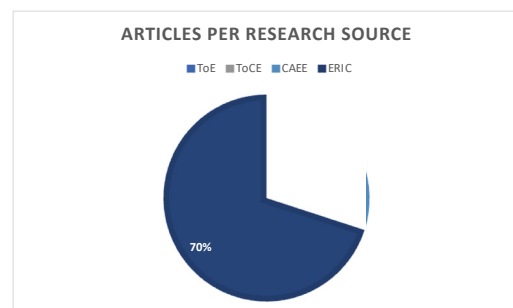


Figure 2: Articles by research source.

It is possible to see that the most significant number of publications occurred in 2019 and 2020, according to Fig. 3.

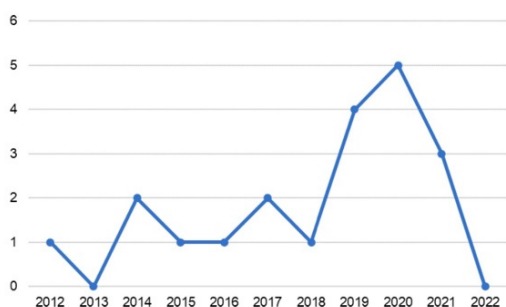


Figure 3: Number of studies on timeline.

This result signals an increase in studies that used AI in education in recent years of research, showing the growing importance of the topic for education. Notably, in the years 2021 and 2022, the number of articles continued to grow. However, some studies from these years were excluded because they focused on assessing the impact of remote teaching during the COVID-19 pandemic. The complete list of selected studies and the details of the selection process are available at <http://bit.ly/3HjtiFX>.

Some limitations are mainly related to the study selection process, which is usually challenging in systematic literature surveys. To mitigate this threat, highly representative research bases in computing education were used in addition to defining quality criteria and applying quality goals based on these criteria. It is also important to emphasize the practical approach of the studies, considered from the definition of the search string, selecting studies that presented proposals and their results in real experiences.

An inherent limitation of RSLs with a qualitative characteristic is the possibility of misinterpretations in selecting and analysing studies related to the research questions. To mitigate the impacts of this characteristic, this study used the strategy advocated by Kitchenham and Charters (2007) of involving more than one person responsible for selecting the studies and a reviewer specialized in the research topic, seeking to guarantee the quality of the analysis.

4 RESULTS

Analysing selected studies, most of them are related to undergraduate students, 47.6% of the total. Meanwhile, schools and graduate courses add up to 19% and 9.5%, respectively. Some studies did not

specify the academic level, equivalent to 23.8%, concerning online courses or virtual learning environments that did not identify the study area, making the objective broader. It is important to note that the total number of studies shown in the graph is greater than the number of selected studies because PS06 is related to undergraduate and graduate studies.

The teaching modality of the studies was classified into three categories: face-to-face, virtual, and hybrid. First, the face-to-face teaching modality consisted of most studies, accounting for 65% of the total and formed mainly by university students. Next, with 25%, comes the modality of virtual teaching environments, which can be found in websites, online games, or online courses. The hybrid modality was the one that least represented the selected studies.

Finally, diverse key words were identified in the studies, as shown the word cloud in Fig. 4.



Figure 4: Context of the studies in key words.

From a word cloud, it is possible to observe that most of the studies were inserted in the context of engineering education. Another popular teaching method was using serious games due to the ease of obtaining student data while interacting with the system, programming practices, programming languages (such as SQL), and Massive Open Online Courses (MOOCs). Many of the studies reported only that they were using data from students at a university and, therefore, were classified more generally as “University.”

Considering the four transpositions in (Cope et al., 2021) no evidence was found about ontologies (namability) in the selected studies. According to Cope et al. (2021), pedagogical and domain ontologies allow AI to track student progress and provide feedback at the right time. However, current digital learning environments, even e-learning environments, are not adequately prepared to exploit the naming capabilities of AI. Students' knowledge and learning are still, for the most part, restricted to traditional methods based on summative content tests. In best-case scenarios, educational data mining can provide trend analysis based on massive data collected by the learning platform, which is still not

yet adapted to provide evidence of learning from learner behaviour in the learning environment, such as keystrokes and click streams.

Considering calculability transposition, it was possible to verify the use of fuzzy logic, algorithms such as K Nearest Neighbour (KNN), Bayesian Networks, and prediction models such as Personalized linear multi-regression (PLMR). Fuzzy logic was frequently cited, having been used by studies PS01, PS04, PS16, and PS17. The studies PS02, PS08, PS10, and PS20 used the KNN algorithm, one of the many supervised learning algorithms used in data mining and machine learning.

About measurability, most of the data used came from the interaction of students with educational games. One example is the PS09 study, which, through logs of students' gameplay in the game Raging Skies, evaluated their knowledge in real-time to change the game's difficulty according to each student's performance. Another form of the measurability transposition was related to online courses, as highlighted in PS15, which used these data to create a model for predicting student performance.

Two main types of representability became evident. First, the system returns the student's evaluation, as reported in PS10, which predicts the student's performance at the end of the semester. The other type cited in PS04 uses the student's assessment to provide personalized instruction and indicate the materials they can study.

Regarding research questions, 90% of studies indicated the opportunity to use AI in student assessment, describing how this technology was applied in this context. Only 65% of studies explicitly mentioned any challenges related to the use of AI, not always with a technological focus but considering human and procedural factors as well. It is worth highlighting that all studies had an applied view on the topic under discussion.

5 DISCUSSIONS

As opportunities for using AI in the assessment of computing students, five categories were identified based on the thematic analysis applied to the studies by the researchers, described in Section 5.1. These opportunities are directly related to impacts on the learning process, discussed at the end of this section. From the perspective of challenges, we found little evidence about them, discussed in Section 2.2.

5.1 RQ1. What Are the Opportunities for Using Ai in Assessing Students in Computing?

Assessment of Skills. Different approaches in this category were identified, in general, focused on the type of diagnostic evaluation to understand the student's profile as a learner or professional. The PS01 study, through fuzzy modelling and using data from students' activities, teachers' opinions, and interaction histories, classified the types of students' learning into four categories: (i) Active or Reflective; (ii) Sensory or Intuitive; (iii) Visual or Verbal; (iv) Sequential or Global. By surveying these categories, it is possible to understand the student's learning profile better and act in a personalized way.

The purpose of PS02 was to evaluate students according to their level of employability, classifying them into most likely employable, likely employable, and less likely employable. For this, it used the KNN algorithm, academic data, and skills needed in a job, such as communication skills.

Using the K-Means classification algorithm and educational data from a virtual English classroom as a dataset, PS05 identified and classified students into different groups according to their knowledge, using AI to diagnose students' performance levels per group.

Another example of AI being used to analyse students' learning profiles, PS06 used data from students' eye movements measured while they attended a class. From these data, it was possible to create a model with the Naïve Bayes algorithm, where it is possible to automatically classify the type of student learning with 71% accuracy. In addition, it was also possible to assess the student's concentration levels.

PS09 focused on formative assessment, considering student interactions with the educational game Raging Skies. The student's mastery level is updated in real-time with Bayesian Knowledge Tracing and Dynamic Bayesian Network algorithms.

Another example of a diagnostic evaluation, PS13 formulated a model using the DP-means algorithm based on data from a pre-test on the DeepTutor ITS platform to group students according to their learning level, classifying them into high, medium, and low.

Assessment of Tasks. This category concerns the studies that automatically evaluate students' tasks and projects; therefore, obtaining a correct and fair evaluation of student performance is particularly challenging. PS17 developed a model using the Fuzzy

AHP algorithm to evaluate and interpret students' projects. This case was assessed based on criteria determined by respective experts, where each one received a weight depending on the quality of the work.

PS08, on the other hand, used data mining techniques to analyse the evaluations made on the work of 672 students from 40 different courses, using Random Forest, KNN, and Support Vector Machines algorithms for the analysis. A grade was attributed to each evaluation, making it possible to identify possible correction errors, and repetitive exercises with short answers increased long-term memory on the studied subject, although requiring considerable teacher effort.

PS18 created a system that automatically generates this type of question automatically from the teaching material using the algorithms of BERT, CoreNLP, and natural language processing techniques.

PS20 used deep learning techniques to evaluate students' work without using labels in the data (not supervised), aiming for an accurate automated feedback system and with near-expert quality assessments.

Assessment of Knowledge. Focusing on the continuous diagnostic evaluation, the study PS04 proposes assessing the students' current knowledge to personalize the teaching. The study presents eLGuide, an ITS that can be used in any e-learning environment. It uses the student's interaction with the system so that, with Fuzzy Modelling, it can understand the student's current knowledge and provide feedback that will help them complete the course satisfactorily.

Prediction of Performance. Focusing on the type of diagnostic evaluation, predicting how the performance of students should be at the end of the term or academic year while the course is beginning, can be very important since, when identifying a student with a high risk of failure, it is possible to take measures to prevent this from happening. PS03 proposed a methodology that classified students' behaviour according to their daily or occasional activities. For this, it used information obtained through RFID tags present in the students and the Bayesian Belief Network algorithm to predict the students' final performance. PS10 showed that it is possible to predict student performance even with limited data. In this case, academic data such as attendance, access to the virtual learning

environment, and the grades of 23 students in the same class were used. Using the KNN algorithm, they showed that it was possible to predict student performance satisfactorily. PS07 and PS11 proposed the analysis of the student's interaction with games to create models to predict the students' final knowledge level, for example, using Naïve Bayes algorithms and support vector machines (PS07). PS12 created a model using decision trees and K-means clustering algorithms to estimate the students' results and satisfaction with the course. Evaluation records, demographic data, and satisfaction surveys were used to calibrate the model. The PS14 created an evaluation model to monitor the learning level during the course based on Natural language processing technologies, such as Word2Vec. PS15 and PS16 used Moodle (a learning management system) to obtain information about students and used models to predict student performance in online courses. PS15 used a personalized regression model, while PS16 used the FRBCS-CHI algorithm (Fuzzy Rule-Based Classification System using Chi's technique) to predict student performance in the first quarter.

Assessment of Questions in Content Tests.

Concerned about the quality of assessment questions, PS19 proposed a system, LosMonitor, to help teachers analyse and monitor the cognitive level of questions. For this, the support vector machine algorithm was used in a dataset with 1630 questions together with the syllabus of 122 courses. The tool classifies questions according to Bloom's Taxonomy and notifies the professor when a question is not aligned with the course objectives. In addition, it presents graphs and statistics to monitor the quality and cognitive level of the questions.

According to the studies analysed, evidence was found that opportunities to use AI have good impacts on the learning process, especially concerning personalizing education, improving student performance, and increasing the quality of the assessment process.

Personalized teaching can make a difference in student performance, providing an environment dedicated to their needs. By evaluating and classifying students' learning profiles, studies PS01 and PS06 were able to generate an educational environment that met the specific needs of each student. PS05 successfully achieved the same objective by grouping students with similar knowledge. In addition to having demonstrated that it is possible to obtain the student's level of expertise in real-time during gameplay, PS09 also showed that its models provide insights into the possible learning

trajectories of students. Identifying students at risk of failing was the concern of studies PS10 and PS16. They successfully predicted students' final performance, enabling teachers to guide struggling students better.

The personalization of navigation in virtual education environments brought by PS04 with eLGuide provided better results in students' general learning than those who did not participate in the experiment. The continuous evaluation reported by PS14 obtained highly reliable results. The monitoring carried out by the tool made it possible for constant feedback to be sent to the students to help them maintain higher levels of motivation and better understand their current level of knowledge. PS02 showed that it was possible to increase students' employability during graduation. Its prediction of the probability of a student's admission to a job proved helpful, as this information makes it possible for teaching centres to start improvement courses through e-learning. PS03 achieved its objective with similar success by allowing specialists to take preventive actions when identifying students with difficulties.

Several studies have shown that it was possible to carry out an automated assessment with Artificial Intelligence and, even so, maintain a quality equal to or better than a traditional one in a more straightforward way and reduce the teachers' workload spent on this type of activity. Despite using different assessment models, PS07 made it possible for teachers to know the student's knowledge precisely on a given subject from the interactions in the game and, therefore, eliminated the need for additional tests, facilitating the application of educational games in teaching. PS19, on the other hand, showed that the information in LosMonitor helped professors create questions of higher quality and more aligned with the course syllabus. PS08 showed that it was possible to successfully identify poorly classified evaluations, making it possible to analyse, for example, which departments have the highest number of errors. PS11 showed that using a hybrid evaluation model (single-task and multi-task models), satisfactory results can be achieved in predicting students' competencies, surpassing the individual models. PS12 showed that analysing different sources of information results in a much richer and more profound assessment of student performance, something that teachers could not quickly achieve through exercises. PS15 reported improved performance in assessing the quality of students' assignments in real-time, considering the students' previous work and information such as engagement and use of study materials. In PS17,

using Fuzzy AHP efficiently modelled the ambiguities of human thinking and provided fair and objective evaluations. PS20 showed a system that predicted students' decisions in a Python course while performing a task, allowing feedback whenever necessary and making the assessment more straightforward, faster, and more consistent.

5.2 What Are the Main Challenges of Using AI from the Selected Contexts?

Few studies commented on research challenges, threats, and limitations concerning RQ2. Among pieces of evidence, we classified the challenges into three groups according to following systemic perspectives:

People:

- The lack of participation of some students in opinion polls (PS12).
- Experiences carried out with small classes (PS05).
- Teachers did not always produce quality material, influencing the results (PS08).
- Data samples from students belonging to the same school, which may have biased the result (PS05).

Processes:

- Data preparation is a process that requires time, effort, and specialized people (PS16).
- For the result of a model to have greater generalization capacity, it should have a more significant amount of data, which was not always available (PS13, PS15).
- Decide fairly and assess students' performances without making any errors in the assessment and evaluation process (PS17).
- The complexity of the natural language made the evaluation difficult (PS08).
- The results of the evaluation model were not satisfactory for a small data set (PS13, PS15).
- Challenges on the measurement models used to calibrate student proficiency levels (PS09).

Technology:

- Limited use of network bandwidth (PS02).
- Instability in the results of some algorithms (PS11).
- Difficulty in calibrating algorithm learning (PS09).
- Data cleaning had to be done manually (PS08).
- Database with erroneously classified entries, or even non-existent (PS12, PS16).

- Little data available, which made it impossible to use specific algorithms (PS13, PS15).
- The accuracy of the model used was not satisfactory (PS14).
- Insufficient resources to run the algorithm on larger datasets; Insufficient chosen taxonomy to give full feedback (PS19).

6 CONCLUSIONS

Most of the studies analysed in this research focused on evaluating undergraduate students due to the greater complexity of activities, exercises, and practices. Using AI to support student assessments, teachers can have more information about students and, with a smaller correction load, can dedicate themselves to the teaching and learning process, making interventions. In general, Artificial Intelligence successfully supports assessment processes and maintains results equal to or superior to traditional assessments in several aspects. Consequently, the number of works related to AI in student assessment increases yearly, showing the subject's growing importance in the academic field. Considering how AI is applied, the algorithm most used by the studies was the Fuzzy model, mainly due to its characteristic of explaining uncertainty. It is important to emphasize that intelligent tools do not replace the role of teachers, so they are being used to support them, improving the quality of the teaching-learning process. The most highlighted challenges in the studies are the technology category, related to AI processes, and problems related to the data sample. Due to missing or misclassified data, many studies spend much more time than expected processing the data, sometimes even manually, impacting the breadth and agility of obtaining results.

As future works, this research intends to investigate the assessment models in detail, associating them with specific objectives beyond better understanding the founded challenges to provide guidelines that minimize them.

REFERENCES

- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in education: principles, policy & practice*, 18(1), 5-25.
- Broisin, J.; Venant, R.; Vidal, P. (2017). Lab4CE: a remote laboratory for computer education. In: *International Journal of Artificial Intelligence in Education (IJAIE)*, 27(1), 154-180.
- Computer Science Curricula 2020, (2020). <http://www.acm.org>, last accessed 2022/13/01
- Cope, B.; Kalantzis, M.; Searsmith, D. (2021). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. In: *Educational Philosophy and Theory*, 53(12), 1229-1245.
- Creasy, R. (2018). *The Taming of Education*. doi: 10.1007/978-3-319-62247-7_6
- Feynman, R. P. (1982). Simulating physics with computers. *International Journal of Theoretical Physics*, 21(6-7), 467-488. doi:10.1007/BF02650179
- Gallardo, K. (2021). The importance of assessment literacy: Formative and summative assessment instruments and techniques. *Workgroups eAssessment: Planning, Implementing and Analysing Frameworks*, 3-25.
- Gorin, J. S. (2007). Test construction and diagnostic testing. *Cognitive diagnostic assessment for education: Theory and applications*, 173-201.
- Gulson, K. N., Sellar, S., & Webb, P. T. (2022). *Algorithms of Education: How datafication and artificial intelligence shape policy*. U of Minnesota Press.
- Harlen, W., & James, M. (1997). *Assessment and learning: differences and relationships between formative and summative assessment*. *Assessment in education: Principles, policy & practice*, 4(3), 365-379.
- Harrow, A. W. (2015). Why now is the right time to study quantum computing. arXiv:1501.00011v1 [quant-ph].
- Hoed, R. M. Dropout analysis in higher education courses (Análise da evasão em cursos superiores). In: *Dissertation (master's dissertation)*, 188 (2016).
- Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. *Cognitive diagnostic assessment for education: Theory and applications*, 19-60.
- Huhta, A. (2008). Diagnostic and formative assessment. *The handbook of educational linguistics*, 469-482.
- Hull, A.; Du Boulay, B. Motivational and metacognitive feedback in SQL- Tutor. *Computer Science Education*, 25(2), 238-256 (2015).
- Kemp, S., & Scaife, J. (2012). Misunderstood and neglected? Diagnostic and formative assessment practices of lecturers. *Journal of Education for Teaching*, 38(2), 181-192.
- Kitchenham, B.; Charters, S. Guidelines for performing systematic literature reviews in software engineering (2007).
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou & K. Q. Weinberger (Eds.) *Neural Information Processing Systems 2012* (pp. 1097-1105).
- Loras, M., Sindre, G., Tr etteberg, H., Aalberg, T. Study behavior in computing education - A systematic literature review. In: *ACM TOCE*, 22(1), 1-40 (2021).
- Lopes, G. B., & dos Santos, S. C. (2021, October). Student Assessment in PBL-Based Teaching Computing: Proposals and Results. In *2021 IEEE Frontiers in Education Conference (FIE)* (pp. 1-9). IEEE.

Maia, D., & dos Santos, S. C. (2022, October). Monitoring Students' Professional Competencies in PBL: A Proposal Founded on Constructive Alignment and Supported by AI Technologies. In 2022 IEEE Frontiers in Education Conference (FIE) (pp. 1-8). IEEE.

Malik, A., Wu, M., Vsavada, V., Song, J., Coats, M., Mitchell, J., Goodman, N.; Piech, C. Generative Grading: Near Human-level Accuracy for Automated Feedback on Richly Structured Problems. In: preprint arXiv:1905.09916 (2019).

Mazza, R.; Milani, C. Exploring usage analysis in learning systems: Gaining insights from visualisations. In: AIED'05 workshop on Usage analysis in learning systems. 65-72 (2005).

Mccarthy, J. What is artificial intelligence? Machine Learning Services on AWS (2007).

Montebello, M. (2019). The ambient intelligent classroom: Beyond the indispensable educator. Dortmund NL: Springer.

Mostafavi, B.; Barnes, T. Evolution of an intelligent deductive logic tutor using data-driven elements. In: International Journal of Artificial Intelligence in Education, 27(1), 5-36 (2017).

Novak, J. D. (2010). Learning, creating and using knowledge: Concept maps as facilitative tools in schools and corporations. New York: Routledge.

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.

Quaye, S. J.; Harper, S. R.; Pendakur. Student engagement in higher education: Theoretical perspectives and practical approaches for diverse populations. In: Routledge (2019).

Rivers, K.; Koedinger, K. R. Data-driven hint generation in vast solution spaces: a self-improving python programming tutor. In: IJAIE 27(1), 37-64 (2017).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536. doi:10.1038/323533a0

Salem, A. B. M. Intellectual e-learning systems. In: International Conference on Virtual and Augmented Reality in Education, 16-23 (2011).

Víctor González-Calatayud, V., Prendes-Espinosa, P., Roig-Vila, R. Artificial Intelligence for Student Assessment: A Systematic Review. In: Appl. Sci., 11(12), 5467 (2021).

Vihavainen, A., Paksula, M., Luukkainen, M. Extreme apprenticeship method in teaching programming for beginners. In: ACM Computer science education, 93-98 (2011).

Yadav, A.; Gretter, S.; Hambrush, S.; Sands, P. Expanding computer science education in schools: understanding teacher experiences and challenges. Computer Science Education, 26(4), 235-254 (2016).

William, D. (2000, November). Integrating formative and summative functions of assessment. In Working group (Vol. 10).

Zafar, A.; Albidewi, I. Evaluation study of eLGuide: A framework for adaptive e-learning. CAEE, 23(4), 542-555 (2015).

Zhai, C., & Massung, S. (2016). Text data management and analysis: A practical introduction to information retrieval and text mining. Williston, VT: ACM and Morgan & Claypool.

Zhang, C., Yang, Z., He, X., & Deng, L. (2020). Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing, 14(3), 478-493.

APPENDIX A

Table 1: Selected studies.

ID	Authors	Title	Year
PS01	Li, N.; Chen, X.; Subramani, S.; Kadry, S.N.	Improved fuzzy-assisted multimedia-assistive technology for engineering education	2020
PS02	Sood, S. K.; Singh, K. D.	Optical fog-assisted smart learning framework to enhance students' employability in engineering education	2019
PS03	Verma, P.; Sood, S. K.; Kalra, S.	Smart computing based student performance evaluation framework for engineering education	2017
PS04	Zafar, A.; Albidewi, I.	Evaluation Study of eLGuide: A Framework for Adaptive e-Learning	2015
PS05	De Moraes, A. M.; Araújo, J. M. F. R.; Costa, E. B.	Monitoring Student Performance Using Data Clustering and Predictive Modelling	2014
PS06	Pritalia, G. L.; Wibirama, S.; Adji, T. B.; Kusrohmaniah, S.	Classification of Learning Styles in Multimedia Learning Using Eye-Tracking and Machine Learning	2020
PS07	Alonso-Fernández, C.; Martínez-Ortiz, I.; Caballero, R.; Freire, M.; Fernández-Manjón, B.	Predicting students' knowledge after playing a serious game based on learning analytics data: A case study	2019

Table 1: Selected studies (cont).

ID	Authors	Title	Year
PS08	Cook, J.; Chen, C.; Reid-Griffin, A.	Using Text Mining and Data Mining Techniques for Applied Learning Assessment	2019
PS09	Cui, Y.; Chu, M. W.; Chen, F.	Analyzing Student Process Data in Game-Based Assessments with Bayesian Knowledge Tracing and Dynamic Bayesian Networks	2019
PS10	Wakelam, E.; Jefferies, A.; Davey, N.; Sun, Y.	The potential for student performance prediction in small cohorts with minimal available attributes	2020
PS11	Henderson, N.; Kumaran, V.; Min, W.; Mott, B.; Wu, Z.; Boulden, D.; Lord, T.; Frieda Reichsman, F.; Dorsey, C.; Wiebe, E.; James Lester, J.	Enhancing Student Competency Models for Game-Based Learning with a Hybrid Stealth Assessment Framework	2020
PS12	Hung, J. L.; Hsu, Y. C.; Rice, K.	Smart computing based student performance evaluation framework for engineering education	2012
PS13	Khayi, N. A.; Rus, V.	Clustering Students Based on Their Prior Knowledge	2019
PS14	Luo, J.; Sorour, S. E.; Goda, K.; Mine, T.	Predicting Student Grade based on Freestyle Comments using Word2Vec and ANN by Considering Prediction Results Obtained in Consecutive Lessons	2015
PS15	Ren, Z.; Rangwala, H.; Johri, A.	Predicting Performance on MOOC Assessments using Multi-Regression Models	2016

ID	Authors	Title	Year
PS16	Zhao, Q.; Wang, J. L.; Pao, T. L.; Wang, L. Y.	Modified Fuzzy RuleBased Classification System for Early Warning of Student Learning	2020
PS17	Çebi, A.; Karal, H.	An application of fuzzy analytic hierarchy process (FAHP) for evaluating students' project	2017
PS18	Lu, O. H. T.; Huang, A. Y. Q.; Tsai, D. C. L.; Yang, S. J. H.	Expert-Authored and Machine-Generated Short-Answer Questions for	2021
PS19	Allamary, A. S.	LOsMonitor: A Machine Learning Tool for Analyzing and Monitoring Cognitive Levels of Assessment Questions	2021
PS20	Malik, A.; Wu, M.; Vasavada, V.; Song, J.; Coots, M.; Mitchell, J.; Goodman, N.; Piech, C.	Generative Grading: Near Human-level Accuracy for Automated Feedback on Richly Structured Problems	2021