Critical Factors for a Reliable AI in Tutoring Systems on Accuracy, Effectiveness, and Responsibility

Davi M. Maia
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
djmm@cin.ufpe.br

Simone C. dos Santos
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
scs@cin.ufpe.br

Luis Gabriel Lima
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
luis.gabriellx@gmail.com

Vinícius Luiz Franca
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
viniciusluiz14052000@gmail.com

Alexsandro Henrique Lima
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
alexsandro.henrique0702@gmail.com

Daniel Andrade
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
daniel.mandrade24@gmail.com

***Abstract*— In recent years, there has been a surge in the development and use of artificial intelligence (AI) systems in various fields, including education. One such application is the AI-based tutoring system, which can provide personalized learning experiences to students. These systems leverage advanced algorithms to analyze student performance, identify knowledge gaps, and deliver targeted feedback and guidance. One of the significant challenges educators and researchers face in the context of AI-based tutoring systems is the lack of reliability in the systems. The accuracy, effectiveness, and responsibility of AI systems are critical factors that determine their reliability. Accuracy challenges for AI algorithms in tutoring systems include accurately modeling individual learner profiles, providing tailored content that aligns with each student's pace and understanding, addressing diverse learning strategies, and ensuring the feedback is specific and actionable. Overcoming data sparsity and ensuring algorithmic transparency and fairness are also significant hurdles. Challenges in ensuring effectiveness include developing algorithms that accurately adapt to individual learning needs, processing natural language effectively, maintaining engagement, and providing contextually relevant feedback. Responsibility challenges include ensuring data privacy and security, preventing algorithm biases affecting learning outcomes, and maintaining ethical standards in AI interactions. Balancing automation with human oversight to support diverse learning needs without compromising educational integrity is also crucial. Considering these challenges, this study discusses the critical factors for reliable AI in tutoring systems from perspectives of accuracy, effectiveness, and responsibility. The research has a descriptive character and qualitative analysis, applying the systematic literature review (SLR) method. The analysis of 43 studies in the last five years made it possible to find some interesting results. In summary, the accuracy of AI in Tutoring Systems is impacted by data quality and preprocessing, choice of appropriate metrics, advanced learning techniques, dataset diversity, and correct selection of hyperparameters. As for the factors that influence the effectiveness of these systems, the personalization of learning and the ability of the systems to adapt to individual needs through personalized pedagogical interventions stand out, using techniques such as recurrent neural networks to predict the quality of interactions. However, challenges related to understanding learning emotions reinforce the complexity of building effective models based on emotion induction. Concerning the responsible use of AI in tutoring systems, it is crucial to consider the privacy and security of student data, adopt collaborative human-machine approaches, and align the use of AI with institutional governance, promoting a safe and ethical learning environment. As a main contribution, we highlight an enlightening discussion of the critical factors for a reliable AI in the context of tutoring systems, identifying quality studies on the subject to support researchers.**

Keywords—Intelligent tutoring system, Critical Factors, Accuracy, Effectiveness, Responsibility.

I. INTRODUCTION

The use of Artificial Intelligence (AI) in education has gained significant attention in recent years. AI-based tutoring systems have emerged as a promising tool to improve student learning outcomes. However, the reliability of these systems is a significant concern, as they must provide accurate, effective, and responsible feedback to students to support their learning process [1].

Tutoring systems have been developed and researched for several decades [1][2][3]. However, the potential of AI in education has gained significant attention in recent years, mainly with the advancement of machine learning algorithms and natural language processing [4][5][6]. The increased availability of data and computational power has also accelerated the development of AI-based tutoring systems, making them more accessible and cost-effective. The COVID-19 pandemic has also highlighted the potential of tutoring systems, as they offer a viable solution to address the challenges of remote learning [7][8][9].

Therefore, it is important to have critical factors that ensure their reliability, as this ensures that students receive realistic and meaningful feedback. Without some critical factors, systems can be inaccurate, ineffective, and ethically questionable, harming students' education and putting their data at risk [1][10][11].

These factors that need to be taken into consideration when building an algorithm for a tutoring system refer to accuracy, effectiveness, and responsibility [2][12][13]. These three factors are essential elements within a project for the

success of the main objective for which that system was built [2][12][13].

Given this reality, the present study's main motivation is to understand if tutoring systems consider aspects such as critical factors for building reliable artificial intelligence. From this objective, the following central research question was defined: "*What is the set of critical factors necessary to ensure the reliability of an artificial intelligence Tutoring System, considering the aspects of accuracy, effectiveness, and responsibility?*" To answer the central question, the following secondary questions were defined: Q1) How is AI used in the context of Tutoring Systems? Q2) What factors impact the accuracy of AI in the context of Tutoring Systems? Q3) What factors influence the effectiveness in the context of Tutoring Systems? Q4) What factors should be considered for the responsible use of AI in the context of Tutoring Systems?

This research used the Systematic Literature Review (SLR) method proposed by Kitchenham [18] in three stages: 1) planning the review; 2) conducting the review and; 3) discussion of the studies. Thus, this paper follows these stages to present its results.

This paper is organized into six sections. After this introduction, Sections II describe a brief conceptual references. Section III describe the related works. After that, in Section IV, presents the research methodology. Section V presents the results. Finally, Section VI presents the conclusions and future perspectives of this research

II. BACKGROUND

According to [1], Intelligent Tutoring Systems (ITS) are computerized learning environments that incorporate computational models from cognitive science, learning, artificial intelligence, and computational linguistics.

These systems track the student's learning progress, creating a model that assesses the level of technical and non-technical knowledge, motivations, and emotions. In general, the interaction starts in a simple and flexible manner to understand the student and their profile. After a few interactions, the system adapts its activities according to the student's needs.

Within the construction of ITS, there are several factors that can contribute to the success of this application. Factors such as accuracy, effectiveness, and responsibility help these systems operate more swiftly and with greater responsiveness, improving interaction with users and achieving the goal [1].

Accuracy in the context of ITS can be understood in two ways. First, as the measure of accuracy of a judgment about a specific task. Second, as the measurement of the relationship between various judgments and corresponding tasks [14,15]. The Pearson correlation coefficient or a contingency coefficient is typically used to measure the relative accuracy of judgments. According to [14,15], immediate feedback is a way to better calibrate the accuracy of ITS judgments regarding which learning activities to recommend.

Effectiveness in the context of ITS can be understood as a set of factors that help the system achieve the goal. According to Van Lenh [2], the effectiveness of an ITS is related to attributes such as: detailed diagnostic assessment, correct selection of individualized activities, tutoring techniques, dialogue control, motivation, feedback, and scaffolding. It is important to emphasize the difficulty in defining an evaluation/construction model, since it is related to a multidimensional concept. Thus, each ITS implements the dimensions that suit its needs.

Responsibility in the context of ITS is related to transparency and security. It is the system's ability to store the information entered by its users without making it available to third parties [16]. Furthermore, it is related to the clear description of the data that will be collected and stored by the system, and the communication of this information to users or their guardians. Finally, responsibility is related to the conscientious use of user data and collecting it solely for improving interaction with them.

III. RELATED WORK

Looking for solutions to the challenges mentioned in Section I, we found some works related to this study.

Tooee is a system that provides text-based visual blocks enabling communication with WebSocket servers, developed by Park and Shin [PS30]. These servers allow Scratch to communicate with text-based programming languages, such as Python and JavaScript. By greatly increasing Scratch's extensibility, students can implement various types of big data/artificial intelligence programs based on the new blocks.

Nieto et al. [PS17] compare three supervised Machine Learning (ML) algorithms that, when used as predictors, improved decision-making at the strategic level. Specifically, they applied Decision Trees, Random Forests, and Logistic Regression to predict graduation rates using real data from a university in South America. ROC curve analysis and Precision are performed as effectiveness measures to compare and evaluate the three algorithms.

Samin and Azim [PS18] highlight the success of the deep learning paradigm; several deep learning-based approaches have also been introduced in combination with collaborative filtering techniques to design recommendation systems. Meanwhile, Ingavélez-Guerra et al. [PS2] address that accessibility and adaptability in digital educational resources constantly require the search for updated research that responds to trends in student learning variability and diversity, demonstrating that the virtual environment is currently considered the most widely used tool in education.

Still focusing on the educational environment, Yanes et al. [PS23] construct a teaching strategy approach based on a predefined set of expected outcomes called OBE. This approach is applied in higher education specifying outcomes at three levels: (1) PEOs, which are general statements describing the career the program is preparing students to achieve, (2) POs, which are more specific statements describing the knowledge, skills, and competencies students are expected to know and be able to do upon graduation, (3) COs, which are statements describing what students are expected to know, the attitudes they are expected to have, and what they are able to do after completing a course.

IV. RESEARCH METHOD

This study used a method based on Kitchenham [17], as shown in Fig. 1. It consists of three main stages, being: 1) planning the review, identifying the motivation, and defining the protocol to be followed; 2) conducting the review, following the defined protocol, extracting, and synthesizing

the data; 3) discussion of the studies, based on the answers found for the research questions.

The final objective is to provide a complete and reliable synthesis of available evidence to guide decision-making and/or identify gaps in the literature.

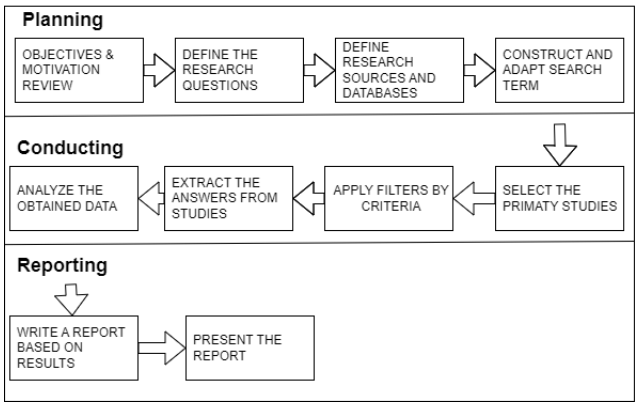

Fig. 1. SLR Process

A. Planning

We aim to answer a primary question: "*What is the set of critical factors necessary to ensure the reliability of an artificial intelligence Tutoring System, considering the aspects of accuracy, effectiveness, and responsibility?*"

To ensure the review addresses the central question, we seek to answer four secondary research questions that defined to guide the search and selection processes:

- Q1) *How is AI used in the context of Tutoring Systems*?
- Q2) *What factors impact the accuracy of AI in the context of Tutoring Systems*?
- Q3) *What factors influence the effectiveness in the context of Tutoring Systems*?
- Q4) *What factors should be considered for the responsible use of AI in the context of Tutoring Systems*?

To answer these research questions, a generic search string was defined based on [17], using search terms focused on the three critical factors and the term tutoring system. We used four research databases: ACM, IEEExplore, Scopus and Science Direct. After several refinements, two different search strings were defined, considering the characteristics of the search tools, as shown in Tables I and II.

TABLE I. ACM & SCOPUS STRING

Search String for ACM & Scopus
("Artificial Intelligence" OR "AI" OR "Machine Learning") AND ("Tutoring Systems" OR "Educational Systems") AND ("Accuracy" OR "Precision") AND ("Efficacy" OR "Effective ness" OR "Effect") AND ("Responsibility" OR "Conscientious")

TABLE II. IEEE & SCIENCE DIRECT STRING

Search String for IEEE & Science Direct
("Artificial Intelligence" OR "AI" OR "Machine Learning") AND ("Tutoring Systems" OR "Educational Systems") AND ("Accuracy" OR "Efficacy" OR "Effectiveness" OR "Responsibility" OR "Conscientious" OR "Precision")

B. Conducting

Using the constructed search string, automatic searches conducted in the digital libraries mentioned earlier. Subsequently, selections made based on their importance and relevance to the primary study area, covering both specialized (ACM and IEEE) and more generalist databases (Scopus and Science Direct). Following the search, filters applied to identify the most relevant works for the present study. The initial filters applied through exclusion criteria based on reading the titles, abstracts, and structures of the works. Exclusion criteria used to select studies consolidated in the selected databases, such as:

- E001 - Content deviating from the focus of the study.
- E002 - Studies not in English or Portuguese.
- E003 - Articles less than 6 pages.
- E004 - Articles longer than 35 pages.
- E005 - Articles that are other systematic mappings of literature.

In the filtering, inclusion criteria for the works also used:

- I001 - Study is available.
- I002 - Study contextualizing a real problem.
- I003 - Complete article.

The above criteria were applied in two distinct stages. First, in filter 1, composed of the information contained in the titles and abstract; later, in filter 2, in case of doubt in selection, considering the introduction and conclusions of the articles for a grounded evaluation. The data collection phase was completed in the middle of 2023.

C. Reporting

After the analysis step, which the articles underwent rigorous inclusion, exclusion, and quality assessment, they proceeded to the final list that address the research questions.

To extract the data, a shared spreadsheet was created containing columns such as title, year of publication, source, authors, and important fragments of the articles' text were extracted and classified. These fragments identified the answer to each research question. At this stage, the specific questions were divided among four team members, undergraduate students in the Information Systems course, each assigned specific activities and supervised by a faculty advisor. The answers from the selected studies were classified according to the relevance of the information to each question. To assess the quality of the studies, criteria such as relevance to the central theme, completeness, and clarity of content, and discussion related to the research questions were adopted. These criteria were evaluated on a three-point scale: 0 - Does not meet; 0.5 - Partially meets; 1 - Fully meets, as shown in

Table III. Each article undergoes an overall quality assessment up to maximum 5 points. To conduct this research stage, four senior students from the Information Systems undergraduate course participated, working in pairs, as well as a Computed Science PhD student and a professor with a doctorate in Computer Science.

TABLE III. QUALITY CRITERIA

ID	Description
1	Clear context
2	Defined methodology
3	Practical application
4	Relevant and consistent discussions
5	Well presented proposal

As a threshold metric for assessing and qualifying studies, a percentage equal to or greater than 50% of the maximum quality grade was defined according to the specified criteria (Clear context, Defined methodology, Practical Application, Relevant and consistent discussions, Well-presented proposal), totalizing 5 points. Thus, studies were classified with a qualification grade greater than or equal to 2.5 points.

D. Limitations and Threats

Considering the literature review, we found some limitations in conducting the research and in the results. First, the experiences presented did not lead to a mapping of critical factors in the context of technology adoption. Thus, data collection was based on the authors' experience and in the experience reports, described in the studies. Another limitation of the search process was the elaboration of the correct string for refinement and adaptation for each database. Since each database adopts a writing style and manages search terms differently, this task requires time and a certain level of knowledge about the area of study and the database usage. It is important to mention that the study selection work was done cautiously due to the difficulty of finding known bases that contain information necessary for SLR.

V. RESULTS

As described earlier, the final selection of articles rigorously followed the research method outlined in Section IV.

First, a breadth-first search was conducted using the Search String in each of the databases, resulting in a total of 82 articles. After this task, we applied the exclusion criteria to the titles and abstracts of each study, resulting in a total of 31 excluded studies and 51 studies accepted.

A second filtering process was conducted, with an in-depth analysis of the introduction and conclusion sections of the studies. This step was important to decide on the inclusion of articles based on the initial filtering. After this step a total of 3 studies was excluded and 48 accepted.

Finally, the articles were classified according to the quality criteria presented in the previous section. At the end of the process, 5 studies were excluded, and 43 studies remained. Fig. 2 summarizes the article selection process according to the PRISMA model.

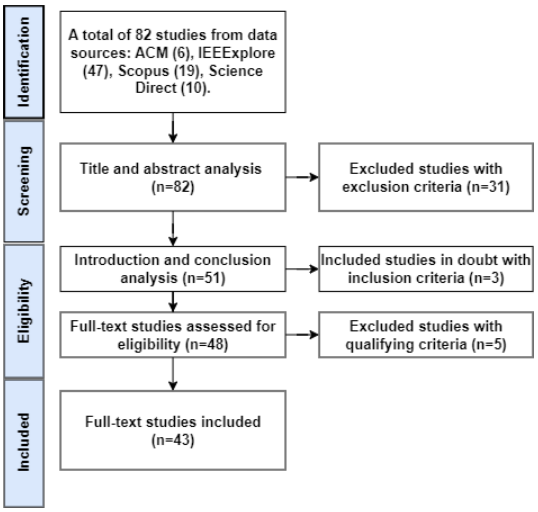

Fig. 2. PRISMA flow chart of section process.

The studies came from different databases, namely: IEEE, Scopus, ACM, and Science Direct, as shown in Fig. 3. The IEEE source stands out with the highest number of relevant works (18/43), followed by Scopus (11/43), Science Direct (8/43), and ACM with only 6 selected studies.

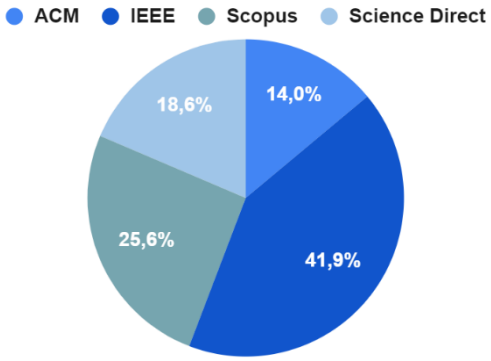

Fig. 3. Source of Studies

Regarding the type of study, the results were balanced, with 28% (12/43) of studies of the conference paper/proceedings type, 62.8% (27/43) of the journals/periodicals type, and 9.2% (4/43) of the chapter type, as shown in Fig 4.

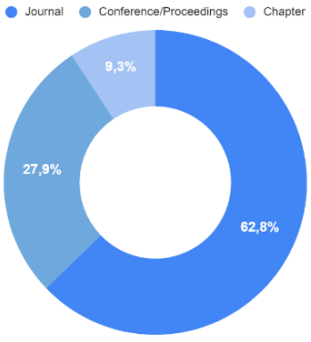

Fig. 4. Type of Studies.

Fig. 5 shows the distribution of studies over time. It can be observed that the year 2022 recorded the highest number of studies on the subject in question.

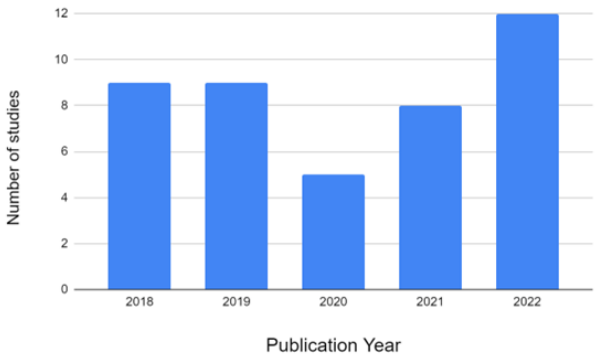

Fig. 5. Number of studies in the timeline.

Regarding the timeline of study production, we found that between 2020 and 2021, there was a reduction in the number of studies found, with 2020 (the year of the COVID-19 pandemic) presenting the lowest number of studies in the last five years. Over the years, there has been a certain consistency in the number of studies, with a significant increase in 2022, when studies on artificial intelligence gained strength due to the rise of generative AI-based technologies such as Chat-GPT.

The following sections discuss, from the perspective of each research question, the findings.

A. *How is AI used in the context of Tutoring Systems?*

After the analysis step, which the articles underwent rigorous inclusion, exclusion, and quality assessment, they proceeded to the final list that address the research questions.

Based on the analysis of the cataloged studies, answers to this question were found in 36 out of 43 studies. As a result, it was possible to identify how artificial intelligence is used in the context of Tutoring Systems.

For example, [PS14] discusses a LEGO-based tutoring system developed to help children with Autism Spectrum Disorder (ASD) improve their communication and social interaction skills. The system is designed to be highly interactive and engaging, using advanced artificial intelligence technologies to personalize each child's learning experience based on their individual needs and abilities. It operates through a user-friendly interface that allows children to interact with a variety of LEGO-based activities. The exercises are designed to help children develop skills such as recognizing facial expressions, understanding emotions, and communication skills. The system uses machine learning techniques to monitor the child's performance in real time, adjusting the difficulty and content of the exercises according to the child's progress.

On the other hand, [PS18] focuses on the higher education scenario in Pakistan. This study presents a recommendation system to help students make informed decisions about their academic careers by implementing machine learning tools for the development of a recommendation system useful for teachers and students in academia. A strength is that the proposed methodology is scalable for different disciplines and can easily adapt to new market trends, enhancing its 'train model' with the latest research publications. Results generated by two popular probabilistic models, LDA and ATM, on real-world data were shown, concluding that LDA's generative performance is much better than ATM, however, ATM provides semantically more useful information than LDA and proves to be more suitable for the recommendation task at hand.

Another use of artificial intelligence was shown in [PS19], which proposed a new individualized AI tutor to help a student achieve a high level of academic success. To consider the student's current status and preferences, they developed the assistant as a system that integrates three DLNs. The proposed tutor was trained with 800,000 training sets collected from a marketed Korean language teaching app. A challenge is to continuously monitor the student's status and preferences and recommend appropriate educational services.

Finally, another example is [PS1], which shows a very comprehensive and futuristic use of AI, which is nothing more than an intelligent classroom system that uses artificial intelligence at various stages, serving as an educational platform that uses advanced artificial intelligence technologies to provide a personalized and adaptive teaching experience to students. It is a future system that contains technological and operational components of an emotionally conscious AI smart classroom that provides automated real-time feedback through two modalities of an open learning model for a presenter during a presentation, in order to improve the effectiveness of the presentation, the presenter's self-regulation and metacognitive awareness, and their verbal and non-verbal communication skills.

Main findings: *These examples demonstrate the diverse applications of AI in Tutoring Systems, ranging from personalized learning for specific student populations to recommendation systems for academic and career guidance and the creation of intelligent classroom environments for enhanced teaching and learning experiences.*

B. *What factors impact AI accuracy in the context of Tutoring Systems?*

Through the analysis of the same studies, answers to this question were found in 18 out of the 43 selected studies, almost half the number of questions answered by question number one. Consequently, it was possible to identify the factors that impact the accuracy of artificial intelligence in the context of Tutoring Systems.

For example, [PS17] discusses the use of machine learning in higher education institutions, which with these experiments allowed the early identification of students with a probability of not graduating, which is highly effective, although from this prediction other aspects, such as students' academic performance and dropout rates, can be analyzed. It highlights the importance of accuracy to ensure that decisions made or results obtained by the system are reliable and accurate, and with this, it guarantees 83% accuracy in the algorithm used for the above finality. The study demonstrates the need to have these students recognized early, as it may allow the governance strategic planning skills of HEIs to

respect student exclusion policies, student dropout rates, retention rates, strengthen programs, and a host of others. To improve accuracy, it mentions that ROC comparison, accuracy, precision rates, and recall rates were performed. Each metric provides different information about the model's accuracy in different areas, which together can obtain a more complete view of the model's accuracy in different areas. This allows developers to adjust the model to improve accuracy in specific areas.

Unlike the primary study shown in the previous paragraph, [PS35] proposed the XAI model to facilitate and help instructors interpret online student behavior studies. The main objective of this research was to make ML models easy to understand in a human-readable way. It points out new essential metrics for improving model accuracy, such as recall, F-score, and confusion matrices. F-score is a measure that combines precision and recall to provide an overall measure of the model's performance in classification problems, while recall measures the proportion of positive instances correctly identified by the model relative to the total positive instances. Finally, the confusion matrix enables the visualization and calculation of several other metrics that assess the quality of XAI and identify adjustments for improving the accuracy percentage.

Furthermore, in [PS5], a new model for diagnosing student knowledge in Intelligent Tutoring Systems (ITS) is proposed. The model uses deep learning techniques to model the interaction between the student's prior knowledge and the information he receives during the learning session. It is based on a recurrent neural network, which is able to capture the sequence of actions taken by the student during the learning session. Therefore, as it aims to diagnose the student's knowledge regarding each learning objective, it is necessary to have a well-evaluated accuracy with a pleasant score. For this, it shows the use of advanced machine learning techniques, such as recurrent neural networks and attention, to model the interaction between the student's prior knowledge and the information he receives during the learning session. These techniques allow the model to more accurately capture the information that the student presents during the learning session, as well as to value the quality of the data, the size of the data set, the appropriate choice of modeling techniques, and the selection of appropriate hyperparameters.

Finally, [PS21] explains about ITS, which are adaptive systems that use intelligent technologies to personalize learning according to the individual characteristics of the student. Unlike most other studies, it mentions the importance of preprocessing to improve model accuracy, as this phase involves a set of techniques used to clean raw data into data more suitable for modeling. In the case of the article, the data is preprocessed to remove noise and redundancies, which are irrelevant, inaccurate, or inconsistent information present in the data that can lead to incorrect conclusions. By removing the noise, preprocessing can improve the quality of the data and the accuracy of the models, reducing the impact of this incorrect information.

Main findings: *These results underscore the importance of employing appropriate evaluation metrics, utilizing advanced ML techniques, and preprocessing data to ensure accurate AI applications in Tutoring Systems. This will ultimately lead to more effective personalized learning experiences.*

C. What factors influence effectiveness in the context of Tutoring Systems?

After analyzing the articles, it was found that 17 answers were found, offering important contributions to understanding the factors that influence the effectiveness of tutoring systems.

[PS35] highlights the implementation of Educational Data Mining (EDM) with the aim of helping all stakeholders involved in online learning and, consequently, improving AI effectiveness, where this study stems from the fact that analyzing students' learning behavior is essential as it helps instructors provide personalized learning content, personalized feedback, and assistance at the right time, keeping students on track.

Furthermore, [PS33] introduced a machine learning-based framework, IntelliDaM, which includes components for feature analysis, unsupervised and supervised learning-based mining, and is useful for improving the performance of data mining tasks. In this study, the effectiveness of the framework is evaluated and analyzed in an EDM case study consisting of real data collected at Babeș-Bolyai University, Romania, over three academic years, for a Computer Science discipline. The evaluation of this framework has a significant gain for the continuity of studies since the proposed IntelliDaM framework is easily configurable and can be applied to data mining in various application domains (e.g., bioinformatics, software engineering, medicine, meteorology, etc.).

The approach proposed in [PS38] is related to effectiveness in the context of Tutoring Systems, as it addresses one of the factors that influence effectiveness in this context, personalization. By using student performance data to generate personalized pedagogical interventions, the intelligent tutoring system can adapt to the individual needs of each student and thus improve the effectiveness of the intervention.

Article [PS6] highlights the importance of using Recurrent Neural Networks (RNNs) to predict question quality. This approach can be applied in Tutoring Systems to predict the effectiveness of the system based on the questions asked by students. By training an RNN model to predict the quality of questions, it is possible to identify patterns and trends that may indicate the system's effectiveness in helping students achieve their educational goals.

In contrast to the positive results of the studies described earlier in this section, the EP8 study proposed a taxonomy of measures used to detect learning emotions classified by the widely used classification system to assess emotions in the dimensions of affective valence called the Self-Assessment Manikin (SAM). This study explores challenges from the perspective of learning models and theories, in which they do not clearly understand the relationship and definition between learning behaviors and the induction of internal emotions. It is the lack of corresponding theories that makes it inefficient to construct a learning model based on the induction of emotions.

Main findings: *Educational Data Mining (EDM) analyzes learning behavior to personalize content, feedback, and assistance, enhancing system effectiveness. Machine*

learning frameworks like IntelliDaM improve data mining tasks, bolstering domain effectiveness. Personalization, guided by student performance data, tailors interventions, boosting system effectiveness. Recurrent Neural Networks (RNNs) predict question quality, aiding in assessing system effectiveness based on student inquiries. Understanding learning behaviors and emotions poses hurdles, potentially affecting system effectiveness.

D. What factors should be considered for the responsible use of AI in the context of Tutoring Systems?

The literature review indicated that research question Q4 obtained a total of 13 answers found in the articles' studies. It is important to mention that this research question also had the lowest number of answers among the questions investigated. This suggests that there is a need for more research on the factors that should be considered to ensure the responsible use of Artificial Intelligence (AI) in tutoring systems. This knowledge is crucial to ensure safety and ethics in the development and implementation of these systems.

In order to offer additional support in their learning, article [PS11] describes an active learning approach for detecting student affect in virtual classrooms. In this study, the factors considered for the responsible use of AI are related to student data privacy. When using AI in tutoring systems, it is important to consider how student data is being collected, stored, and used. The use of activity sensors described in the study may be a safer and more private option than other data collection methods, such as video recordings of students' facial expressions. The advantage of this data collection method is that it can be less intrusive and can be anonymized more easily to protect student privacy.

[PS40] highlights the collaborative approach as a means of shifting teachers' focus to help each student, thus providing more personalized support, and thus making the use of AI more responsible by taking the excessive responsibility off the algorithm to provide constant feedback. As the following excerpt highlights, "Teachers' energy can be focused on helping each student, providing more accurate and personalized follow-up. Combined with the needs of collaborative human-machine education and the reality of education and teaching in primary and secondary schools, referring to the architecture and main functions of the existing intelligent learning system and the six characteristics of human-machine collaboration to promote accuracy, collaboration, optimization, personality, thinking, and wisdom creation."

Regarding [PS1] which proposes a smart classroom system that consists of the learning experience and continuous communication between students and teachers using real-time detection and machine intelligence, it highlights its concern for the security and privacy of sensitized data through the use of known encryption-based techniques that can be used to ensure secure (i.e., encrypted) transport of data from mobile devices to the cloud and vice versa.

Lastly, article [PS17] discusses the use of machine learning to support strategic decision-making in higher education institutions. In it, the responsible use of AI is exercised according to institutional governance. The authors seek to understand how institutional governance (i.e., the set of rules, processes, and governance structures that govern the decisions and management of the institution) influences the decision-making structure in Higher Education Institutions (HEIs). In the excerpt, "Continuing the research conducted in a previous work, we compare different Machine Learning algorithms in this article, as well as analyze the decision-making structure in HEIs and how they are managed according to institutional governance," the author is comparing different Machine Learning algorithms to support the decision-making structure in HEIs. They are analyzing how HEIs are managed according to institutional governance, i.e., how decisions are made in the institution and how governance influences this decision-making process.

Main findings: *Responsible AI use entails considerations such as student data privacy, collaborative approaches, security measures, and adherence to institutional governance. The studies emphasize privacy-preserving data collection methods and collaborative teacher-AI interactions for personalized support, the importance of encryption-based techniques for data security in intelligent classroom systems, and the significance of institutional governance in guiding AI applications and decision-making processes in higher education.*

VI. Conclusions

Based on the investigated studies, several articles were found to respond well to the central research question. The review also pointed out the complexity and specificity of the topic, which made it difficult to identify a direct answer to the question. Overall, the studies indicate that the reliability of an AI-based tutoring system depends on several interrelated factors, such as data quality, transparency and interpretability of algorithms, ability to adapt to the individual needs of students, among others. Building reliable tutoring systems requires a multidisciplinary approach and joint effort among AI experts, educators, and researchers in ethics and social responsibility.

As future work, the authors envision a deeper investigation into the development of new algorithms, implementation of personalized systems, ensuring the responsibility and transparency of AI systems, improving inclusion and equity in education, emotionally intelligent and motivational feedback to students, and investigating the ethical and legal implications of ITS use. Significant advances in this field are expected, leading to more effective and inclusive education for all students.

References

[1] Graesser, A. C., Conley, M. W., & Olney, A, “Intelligent tutoring systems”, 2012.

[2] VanLehn, K., “The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems.”, Educational psychologist, 46(4), 197-221, 2011.

[3] Polson, M. C., & Richardson, J. J., “Foundations of intelligent tutoring systems.”,Psychology Press, 2013.

[4] Cerezo Menéndez, R., Esteban García, M., Vallejo Seco, G., Sánchez Santillán, M., & Núñez Pérez, J. C., “Differential efficacy of an intelligent tutoring system for university students: A case study with learning disabilities.”, Sustainability (Switzerland), 2020.

[5] Younis, H. A., Ruhaiyem, N. I. R., Ghaban, W., Gazem, N. A., & Nasser, M., “A Systematic Literature Review on the Applications of Robots and Natural Language Processing in Education.”, Electronics, 12(13), 2864, 2023.

[6] Rathore, A. S., & Arjaria, S. K. “Intelligent tutoring system.”, In Utilizing educational data mining techniques for improved learning: emerging research and opportunities (pp. 121-144). IGI global, 2020.

[7] Mirchi, N., Ledwos, N., & Del Maestro, R. F., “Intelligent tutoring systems: re-envisioning surgical education in response to COVID-19.”, Canadian Journal of Neurological Sciences, 48(2), 198-200, 2021.

[8] Cao, J., Yang, T., Lai, I. K. W., & Wu, J. “RETRACTED: Student acceptance of intelligent tutoring systems during COVID-19: The effect of political influence.”, International Journal of Electrical Engineering & Education, 60(1_suppl), 2495-2509, 2023.

[9] del Olmo-Muñoz, J., González-Calero, J. A., Diago, P. D., Arnau, D., & Arevalillo-Herráez, M. “Intelligent tutoring systems for word problem solving in COVID-19 days: could they have been (part of) the solution?”, ZDM–Mathematics Education, 55(1), 35-48, 2023.

[10] Clancey, W. J., “Intelligent Tutoring Systems: A Tutorial Survey”, 1986.

[11] Yuce, A., Abubakar, A. M., & Ilkan, M., “Intelligent tutoring systems and learning performance: Applying task-technology fit and IS success model.”, Online Information Review, 43(4), 600-616. 2019.

[12] Kulik, J. A., & Fletcher, J. D., “Effectiveness of intelligent tutoring systems: a meta-analytic review.”, Review of educational research, 86(1), 42-78, 2016.

[13] Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q., “Intelligent tutoring systems and learning outcomes: A meta-analysis.”, Journal of educational psychology, 106(4), 901, 2014.

[14] Feyzi-Behnagh, R., Azevedo, R., Legowski, E., Reitmeyer, K., Tseytlin, E., & Crowley, R. S., “Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system.”, Instructional science, 42, 159-181, 2014.

[15] Mousavinasab, E., Zarifsanaiey, N., R. Niakan Kalhori, S., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M., “Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods.”, Interactive Learning Environments, 29(1), 142-163, 2021.

[16] Mougiakou, E., Papadimitriou, S., & Virvou, M., “Intelligent tutoring systems and transparency: The case of children and adolescents.”, In 2018 9th international conference on information, intelligence, systems and applications (IISA) (pp. 1-8), IEEE, 2018.

[17] B. Kitchenham, S. Charters, “Guidelines for performing systematic literature reviews in software engineering,” 2007.

Selected Papers

[PS1] Kim, Y., Soyata, T., and Behnagh R. F., “Towards Emotionally Aware AI Smart Classroom: Current Issues and Direc tions for Engineering and Education”. In: IEEE Access 6, pp. 5308–5331. doi: 10.1109/ ACCESS.2018.2791861. 2018.

[PS2] Ingavélez-Guerra, P., Robles-Bykbaev, V. E., Perez-Muñoz, A., Hilera-González, J., and Otón-Tortosa, S., “Automatic Adaptation of Open Educational Resources: An Approach From a Multilevel Methodology Based on Students’ Preferences, Educational Special Needs, Artificial Intelligence and Accessibility Metadata”. In: IEEE Access 10, pp. 9703– 9716. doi: 10.1109/ACCESS.2021.3139537, 2022.

[PS3] Ramachandran, S., Jensen, R., Ludwig, J., Domeshek, E., and Haines, “T. ITADS: a real-world intelligent tutor to train troubleshooting skills.”, In Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, Proceedings, Part II 19 (pp. 463-468), Springer International Publishing, 2018.

[PS4] Goldberg, B., Roberts, N., Powell, W. G., and Burmester, E., “Intelligent tutoring in the wild: leveraging mobile app technology to guide live training.”, In Proceedings of the Defense and Homeland Security Simulation (DHSS) Workshop, Budapest, Hungary, 2018.

[PS5] Holstein, K., Yu, Z., Sewall, J., Popescu, O., McLaren, B. M., and Aleven, V., “Opening up an intelligent tutoring system development environment for extensible student modeling.”, In Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, Proceedings, Part I 19 (pp. 169-183), Springer International Publishing, 2018.

[PS6] Ruseti, S., Dascalu, M., Johnson, A. M., Balyan, R., Kopp, K. J., McNamara, D. S., and Trausan-Matu, S., “Predicting question quality using recurrent neural networks.”, In Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, Proceedings, Part I 19 (pp. 491-502), Springer International Publishing, 2018.

[PS7] Hagiya, T., Hoashi, K., and Kawahara, T., “Voice input tutoring system for older adults using input stumble detection.”, In 23rd International Conference on Intelligent User Interfaces (pp. 415-419), 2018.

[PS8] Xu, T., Zhou, Y., Wang, Z., and Peng, Y., “Learning emotions EEG-based recognition and brain activity: A survey study on BCI for intelligent tutoring system.”, Procedia computer science, 130, 376-382, 2018.

[PS9] Chen, Z., Salazar, E., Marple, K., Das, S. R., Amin, A., Cheeran, D., and Gupta, G., “An AI-based heart failure treatment adviser system.”, IEEE journal of translational engineering in health and medicine, 6, 1-10, 2018.

[PS10] Cui, W., Xue, Z., and Thai, K. P., “Performance comparison of an AI-based adaptive learning system in China”, In 2018 Chinese automation congress (CAC) (pp. 3170-3175), IEEE, 2018.

[PS11] Yang, T. Y., Baker, R. S., Studer, C., Heffernan, N., and Lan, A. S. “Active learning for student affect detection.”, In Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019. International Educational Data Mining Society (IEDMS) 2019 (pp. 208-217). Université du Québec; Polytechnique Montréal, 2019.

[PS12] Aravind, T., Reddy, B. S., Avinash, S., and Jeyakumar, G. “A comparative study on machine learning algorithms for predicting the placement information of under graduate students.”, In 2019 Third International conference on I-SMAC (IoT in social, mobile, analytics and cloud) (I-SMAC) (pp. 542-546), IEEE, 2019.

[PS13] Kumar, A., Chavan, P., and Mitra, R. “Can EEG signal predict learners’ perceived difficulty?”, In Proc. 27th Int. Conf. Comput. Educ, (Vol. 1, pp. 63-68), 2019.

[PS14] Sun, Q., and Winoto1, P., “An Intelligent LEGO tutoring system for children with autism spectrum disorder.”, In Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence (pp. 56-59), 2019.

[PS15] Llanda, C. J. R., “Video tutoring system with automatic facial expression recognition: an enhancing approach to e-learning environment.”, In Proceedings of the 2019 4th International Conference on Intelligent Information Technology (pp. 5-9), 2019.

[PS16] Georgila, K., Core, M. G., Nye, B. D., Karumbaiah, S., Auerbach, D., and Ram, M., “Using reinforcement learning to optimize the policies of an intelligent tutoring system for interpersonal skills training.”, In Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (pp. 737-745), 2019.

[PS17] Nieto, Y., Gacía-Díaz, V., Montenegro, C., González, C. C., and Crespo, R. G. “Usage of machine learning for strategic decision making at higher educational institutions.”, Ieee Access, 7, 75007-75017. 2019.

[PS18] Samin, H., and Azim, T., “Knowledge based recommender system for academia using machine learning: a case study on higher education landscape of Pakistan.”, IEEE Access, 7, 67081-67093, 2019.

[PS19] Kim, W. H., and Kim, J. H., “Individualized AI tutor based on developmental learning networks.”, IEEE Access, 8, 27927-27937, 2020.

[PS20] Ostrander, A., Bonner, D., Walton, J., Slavina, A., Ouverson, K., Kohl, A., and Winer, E., “Evaluation of an intelligent team tutoring system for a collaborative two-person problem: Surveillance.”, Computers in human behavior, 104, 105873, 2020.

[PS21] Muangprathub, J., Boonjing, V., and Chamnongthai, K. “Learning recommendation with formal concept analysis for intelligent tutoring system.”, Heliyon, 6(10), 2020.

[PS22] Verma, C., Stoffová, V., Illés, Z., Tanwar, S., and Kumar, N. “Machine learning-based student’s native place identification for real-time.”, IEEE Access, 8, 130840-130854, 2020.

[PS23] Yanes, N., Mostafa, A. M., Ezz, M., and Almuayqil, S. N., “A machine learning-based recommender system for improving students learning experiences.”, IEEE Access, 8, 201218-201235, 2020.

[PS24] Tedre, M., Toivonen, T., Kahila, J., Vartiainen, H., Valtonen, T., Jormanainen, I., and Pears, A., “Teaching machine learning in K–

12 classroom: Pedagogical and technological trajectories for artificial intelligence education.", IEEE access, 9, 110558-110572, 2021.

[PS25] Nauman, M., Akhtar, N., Alhudhaif, A., nd Alothaim, A. "Guaranteeing correctness of machine learning based decision making at higher educational institutions." IEEE access, 9, 92864-92880, 2021.

[PS26] Qu, Y., and Ogunkunle, O., "Enhancing the Intelligence of the Adaptive Learning Software through an AI assisted Data Analytics on Students Learning Attributes with Unequal Weight.", In 2021 IEEE Frontiers in Education Conference (FIE) (pp. 1-6), IEEE, 2021.

[PS27] Chaudhary, A., Belani, M., Maheshwari, N., and Parnami, A., "Verbose: Designing a Context-based Educational System for Improving Communicative Expressions", In Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction (pp. 1-13), 2021.

[PS28] Conati, C., Barral, O., Putnam, V., and Rieger, L., "Toward personalized XAI: A case study in intelligent tutoring systems.", Artificial intelligence, 298, 103503, 2021.

[PS29] Alonso-Secades, V., López-Rivero, A. J., Martín-Merino-Acera, M., Ruiz-García, M. J., and Arranz-García, O., "Designing an intelligent virtual educational system to improve the efficiency of primary education in developing countries", Electronics, 11(9), 1487, 2022

[PS30] Park, Y., and Shin, Y., "Tooee: A novel scratch extension for K-12 big data and artificial intelligence education using text-based visual blocks.", IEEE Access, 9, 149630-149646, 2021.

[PS31] Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., and Ghani, N. A. M., "Multiclass prediction model for student grade prediction using machine learning.", IEEE Access, 9, 95608-95621, 2021

[PS32] Roy, R., Babakerkhell, M. D., Mukherjee, S., Pal, D., and Funilkul, S., "Evaluating the intention for the adoption of artificial intelligence-based robots in the university to educate the students.", IEEE Access, 10, 125666-125678, 2022.

[PS33] Czibula, G., Ciubotariu, G., Maier, M. I., and Lisei, H., "IntelliDaM: A machine learning-based framework for enhancing the performance of decision-making processes. A case study for educational data mining.", IEEE Access, 10, 80651-80666, 2022.

[PS34] Zafari, M., Bazargani, J. S., Sadeghi-Niaraki, A., and Choi, S. M., "Artificial intelligence applications in K-12 education: A systematic literature review.", Ieee Access, 10, 61905-61921. 2022.

[PS35] Adnan, M., Uddin, M. I., Khan, E., Alharithi, F. S., Amin, S., and Alzahrani, A. A., "Earliest possible global and local interpretation of students' performance in virtual learning environment by leveraging explainable AI.", IEEE Access, 10, 129843-129864, 2022.

[PS36] Abd El-Haleem, A. M., Eid, M. M., Elmesalawy, M. M., and Hosny, H. A. H. (2022). A generic ai-based technique for assessing student performance in conducting online virtual and remote-controlled laboratories. Ieee Access, 10, 128046-128065.

[PS37] Kim, J., and Shim, J., "Development of an AR-based AI education app for non-majors.", IEEE Access, 10, 14149-14156, 2022.

[PS38] Kochmar, E., Vu, D. D., Belfer, R., Gupta, V., Serban, I. V., and Pineau, J., "Automated data-driven generation of personalized pedagogical interventions in intelligent tutoring systems.", International Journal of Artificial Intelligence in Education, 32(2), 323-349, 2022

[PS39] Koushik, K. S., Chengappa, B. S., and Chendan, R. P., "Automated marks entry processing in Handwritten answer scripts using character recognition techniques.", In 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 728-733), IEEE, 2022.

[PS40] Zhao, X., and Chen, M., "The Functional Architecture Design and Value Orientation of the Intelligent Tutoring System from the Perspective of Man-Machine Collaboration.", In Proceedings of the 2021 4th International Conference on Education Technology Management (pp. 67-72), 2021.

[PS41] Na, W. E. I., Feng, Y. A. N. G., Muthu, B., and Shanthini, A., "Human machine interaction-assisted smart educational system for rural children.", Computers and Electrical Engineering, 99, 107812, 2022.

[PS42] Su, Y., Cheng, Z., Wu, J., Dong, Y., Huang, Z., Wu, L.,and Xie, F. "Graph-based cognitive diagnosis for intelligent tutoring systems.", Knowledge-Based Systems, 253, 109547. 2022.

[PS43] Gan, W., Sun, Y., and Sun, Y. "Knowledge interaction enhanced sequential modeling for interpretable learner knowledge diagnosis in intelligent tutoring systems.", Neurocomputing, 488, 36-53, 2022.