UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

Ph.D. DEGREE IN
Mathematics and Computer Science
Cycle XXXVI

TITLE OF THE Ph.D. THESIS

Characterizing Algorithmic Performance in Machine Learning for Education

Scientific Disciplinary Sector(s)

INF/01

Ph.D. Student:	Roberta Galici
Supervisor	Gianni Fenu
Co-Supervisor	Mirko Marras

Final exam. Academic Year 2022/2023
Thesis defence: February 2024 Session

Statement of Authorship

I declare that this thesis entitled "Characterizing Algorithmic Performance in Machine Learning for Education" and the work presented in it are my own. I confirm that:

- this work was done while in candidature for this PhD degree;
- when I consulted the work published by others, this is always clearly attributed;
- when I quoted the work of others, the source is always given;
- I have acknowledged all main sources of help;
- with the exception of the above references, this thesis is entirely my own work;
- appropriate ethics guidelines were followed to conduct this research;
- for work done jointly with others, my contribution is clearly specified.

Abstract

The integration of artificial intelligence (AI) in educational systems has revolutionized the field of education, offering numerous benefits such as personalized learning, intelligent tutoring, and data-driven insights. However, alongside this progress, concerns have arisen about potential algorithmic disparities and performance issues in AI applications for education. This doctoral thesis addresses these concerns and aims to foster the development of AI in educational contexts that emphasize performance analysis.

The thesis begins by investigating the challenges and needs of the educational community in integrating responsible practices into AI-based educational systems. Through surveys and interviews with experts in the field, real-world needs and common areas for developing more responsible AI in education are identified.

According to our findings, further research delves into the analysis of student behavior in both synchronous and asynchronous learning environments. By examining patterns of student engagement and predicting student success, the thesis uncovers potential performance issues (e.g., unknown unknowns: the model is really confident of its predictions but actually wrong), emphasizing the need for nuanced approaches that consider hidden factors impacting students' learning outcomes.

By providing an integrated view of the performance analyses conducted in different learning environments, the thesis offers a comprehensive understanding of the challenges and opportunities in developing responsible AI applications for education. Ultimately, this doctoral thesis contributes to the advancement of responsible AI in education, offering insights into the complexities of algorithmic disparities and their implications. The research work presented herein serves as a guiding framework for designing and deploying AI-enabled educational systems that prioritize responsibility, and improved learning experiences.

Biography

Roberta Galici was born on *April 12, 1996* in *Cagliari* (Italy). She is a PhD Candidate in Computer Science at the *Department of Mathematics and Computer Science*, *University of Cagliari* (Italy), under the supervision of Prof. *Gianni Fenu* and co-supervision of *Dr. Mirko Marras*. She received the MSc Degree in Computer Science (cum laude) from the same University in 2020.

In 2022, she spent six months at *EPFL* (Switzerland), collaborating with the *ML4ED: Machine Learning for Education* Laboratory. In 2023, she completed a leadership program on *Women's Leadership* at *Bocconi University*. She also achieved *C1 level proficiency* in English in 2021, certified by Cambridge Assessment English at Anglo American.

Since 2021, she has been a teaching assistant for the "*Computer Networks*" course and has assisted three bachelor's degree students with their theses in Computer Science.

She has been awarded a *Globusdoc* grant to spend *three* months abroad.

Her research interests focus on machine learning for educational platforms. He has co-authored papers in top-tier international conferences, such as *AIED*, *LAK*, and *UMAP*. He has given talks and poster presentations at several conferences and workshops, such as *L2D 2021*, *ECSS 2021*, *AIED 2022*, *HELMETO 2022*, *LAK 2023*.

Dissemination

The skills and content covered in this Ph.D. thesis have been shaped by the culmination of research efforts detailed in *6* papers published in national and international conference proceedings. I extend sincere gratitude to my co-authors for their invaluable contributions, which I acknowledge through the inclusive use of the scientific 'we' throughout this thesis. Additionally, during my period abroad in Switzerland, I had the privilege of collaborating with Prof. Tanja Kaser.

It is essential to clarify my individual contributions. I conceived the research concepts outlined in this thesis and undertook the majority of the research work. I conceptualized the methodologies, determined the research trajectories, and collected and analyzed the necessary datasets. The responsibility for script implementation also fell within my purview. Furthermore, I undertook the authorship of the papers, expertly navigating the peer-review process and iteratively refining them. My interactions with the co-authors were characterized by close collaboration and consultation. Their input encompassed offering insights into methodologies, providing technical assistance, engaging in the exploration of techniques, and contributing to the refinement of submitted work. Additionally, I assumed the role of presenter for *4* papers at conferences and workshops.

The detailed references to the produced papers are provided below.

Peer-reviewed Publications in International Conference Proceedings

i. **Roberta Galici**, & Tanja Kaser, Gianni Fenu, and Mirko Marras. (2023). *How Close are Predictive Models to Teachers in Detecting Learners at Risk?*. In Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23). Association for Computing Machinery, New York, NY, USA, 135–145. https://doi.org/10.1145/3565472.3595620 (Rank B)

ii. **Roberta Galici**, & Tanja Kaser, Gianni Fenu, and Mirko Marras. (2023). *Do Not Trust a Model Because It is Confident: Uncovering and Characterizing Unknown Unknowns to Student Success Predictors in Online-Based Learning.*. In LAK23: 13th International Learning Analytics and Knowledge Conference (LAK2023). Association for Computing Machinery, New York, NY, USA, 441–452. https://doi.org/10.1145/3576050.3576148 (Rank A) (Presenter)

iii. Fenu, G., & **Galici, R.**, Marras, M. (2022). *Experts' View on Challenges and*

Needs for Fairness in Artificial Intelligence for Education. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds) Artificial Intelligence in Education. AIED 2022. Lecture Notes in Computer Science, vol 13355. Springer, Cham. https://doi.org/10.1007/978-3-031-11644-5_20 (Rank A) (Presenter)

iv. Fenu, G., & **Galici, R.**, Marras, M., Picciau, S. (2022). *Supporting Instructors with Course Attendance and Quality Prediction in Synchronous Learning.* In: Fulantelli, G., Burgos, D., Casalino, G., Cimitile, M., Lo Bosco, G., Taibi, D. (eds) Higher Education Learning Methodologies and Technologies Online. HELMeTO 2022. Communications in Computer and Information Science, vol 1779. Springer, Cham. https://doi.org/10.1007/978-3-031-29800-4_6 (Rank Unknown)

v. Fenu, G., & **Galici, R.**, Marras, M., Picciau, S. (2022). *Exploiting Student Participation for Early Prediction of Course Quality in Universities.* In: 4th International Conference on Higher Education Learning Methodologies and Technologies Online. HELMeTO 2022. https://shorturl.at/mzY49 (Rank Unknown) (Presenter)

Peer-reviewed Publications in International Workshop Proceedings

i. Fenu, G., & **Galici, R.**, Marras, M. (2021). *Modelling Student Behavior in Synchronous Online Learning during the COVID-19 Pandemic..* L2D@ WSDM. In: L2D'21: First International Workshop on Enabling Data-Driven Decisions from Learning on the Web https://ceur-ws.org/Vol-2876/paper3.pdf (Rank A*) (Presenter)

ii. Fenu, G., & **Galici, R.**, Marras, M. (2021). *Auditing Machine Learning Models for Online Educational Platforms.* In: The 1st Early Career Researchers Workshop Collocated with ECCS 2021 https://shorturl.at/szLXZ (Rank Unknown) (Presenter)

Acknowledgements

I would like to extend my heartfelt gratitude to the individuals who have been pivotal in shaping my academic journey and the successful completion of this doctoral thesis.

First and foremost, I express my sincerest appreciation to my esteemed advisor, Prof. Gianni Fenu, for his guidance, wisdom, and unwavering support throughout this research endeavor. Your insights have illuminated my path and enriched the depth of my work. I am equally indebted to my co-advisor, Dr. Mirko Marras, for his valuable contributions, mentorship, and critical feedback that have been instrumental in refining my research approach. My time at École polytechnique fédérale de Lausanne (EPFL) would not have been possible without the gracious hospitality of Prof. Tanja Kaser. Her guidance, warmth, and the opportunity to collaborate during my 6-month stay have been truly transformative, and I am immensely grateful for the enriching experience. I also extend my heartfelt thanks to my colleagues, friends, and family for their unwavering support, encouragement, and patience throughout this academic journey. Your belief in me has been a constant source of motivation.

I am thankful to the research funding agencies that have provided financial support, enabling me to conduct the research presented in this thesis. In closing, I want to express my deep appreciation to all those who have touched my academic and personal life, contributing to the person I am today.

Nomenclature

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
EDA	Exploratory Data Analysis
EDM	Educational Data Mining
EML	Educational Machine Learning
K-NN	K-Nearest Neighbors
ML	Machine Learning
MOOC	Massive Open Online Course
RF	Random Forest
SVM	Support Vector Machine
UU	Unknown Unknowns

Numerical Expressions

{n}K	{n} thousands
{n}M	{n} millions

Contents

Statement of Authorship **I**

Abstract **II**

Biography **III**

Dissemination **VI**

Acknowledgments **VIII**

Nomenclature **X**

List of Figures **XVII**

List of Tables **XIX**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Challenges . . . 2
1.3 Contributions . . . 2
1.4 Outline . . . 2

2 Background on Machine Learning for Education **5**
2.1 Introduction . . . 5
2.1.1 Role of ML in Education . . . 5
2.1.2 Benefits of ML in Education . . . 6
2.2 Educational Environments . . . 7
2.2.1 Traditional Face-to-Face Education . . . 7
2.2.2 Synchronous Learning . . . 8
2.2.3 Asynchronous Learning . . . 9
2.3 Data Mining Techniques . . . 10
2.4 Data Mining Tools . . . 11
2.5 Data Mining Process . . . 12
2.6 Educational Applications . . . 14

3 Needs and Challenges in Machine Learning for Education 17
3.1 Introduction . . . 17
3.2 Methodology . . . 19
3.2.1 Survey Study Implementation . . . 19
3.2.2 Interview Study Implementation . . . 19
3.2.3 Survey and Interview Data Analysis . . . 21
3.3 Results and Discussion . . . 21
3.3.1 Challenges and needs in data collection and grouping . . . 22
3.3.2 Challenges and needs of fairness-aware technical pipelines . . . 23
3.3.3 Challenges and needs in providing fairness guarantees . . . 24
3.3.4 Challenges and needs of a more holistic fairness auditing . . . 25
3.3.5 Challenges and needs in team blind spots and practices . . . 26
3.4 Findings and Recommendations . . . 27

4 Analyses of Algorithmic Performance in Synchronous Learning 29
4.1 Error Analysis on Student Behavior Modelling . . . 29
4.1.1 Introduction . . . 29
4.1.2 Related Work . . . 30
4.1.3 Methodology . . . 32
4.1.4 Results and Discussion . . . 34
4.1.5 Findings and Recommendations . . . 39
4.2 Error Analysis on Course Attendance Modelling . . . 40
4.2.1 Introduction . . . 40
4.2.2 Methodology . . . 41
4.2.3 Experimental Results . . . 46
4.2.4 Findings and Recommendations . . . 50

5 Analyses of Algorithmic Performance in Asynchronous Learning 53
5.1 Error Analysis on Student Success Prediction . . . 53
5.1.1 Introduction . . . 53
5.1.2 Methodology . . . 56
5.1.3 Experimental Results . . . 64
5.1.4 Findings and Recommendations . . . 70
5.2 Model-Human Comparison on Student Success Prediction . . . 72
5.2.1 Introduction . . . 72
5.2.2 Methodology . . . 73
5.2.3 Experimental Results . . . 79
5.2.4 Findings and Recommendations . . . 87

6 Conclusions 91
6.1 Contribution Summary . . . 91
6.2 Take-home Messages . . . 92
6.3 Future Research Directions . . . 92

List of Figures

1.1 Key Contributions of the Thesis. 3

2.1 Interconnected steps in the data mining process for AI-driven educational systems 13

3.1 Sample population statistics for our survey and interview process. 20

4.1 Overview of Bachelor's and Master's Degree Programmes per Faculty. . 33
4.2 Schematic representation of the online synchronous environment. 34
4.3 Average number of students per lesson for the six faculties. 36
4.4 Centroids of the clusters identified for the nine considered courses. . . . 38
4.5 Methodology overview: Collecting student participation logs, extracting relevant features, and training classifiers for predicting course attendance and quality. 42
4.6 Models performance (AUC) on the course quality prediction task (RQ1). . 47
4.7 Models performance (AUC) on the next lesson participation task (RQ2). . 48
4.8 Percentage of students for whom the course attendance requirement prediction was not trivial (RQ3). 49
4.9 Performance in terms of AUC for the course requirement task (RQ3). . . 49

5.1 Motivating Example for the Unknown Unknowns Problem. 54
5.2 Methodology steps for detecting and characterizing students performance 55
5.3 Student Success Models Performance 59
5.4 Student Success Predicted Probabilities. 61
5.5 Prediction Student Groups: AVG percentage of students in train, val and test. 65
5.6 R2 score of the linear regression models and AVG coefficients 66
5.7 Instructors' Perception. 69
5.8 Methodology steps for human understanding. 74
5.9 Distribution of teacher prediction (RQ1) 80
5.10 Percentage of teacher decision (RQ1) 81
5.11 Distribution of teachers' confidence for their prediction (RQ2) 82
5.12 Percentage of teachers showing a confidence (RQ2) 84

List of Tables

2.1 Data Mining Techniques for Student Performance Prediction. 10

4.1 Example of data recorded for a lesson of a given course. 35
4.2 Schema of the data structure and fields leveraged in this study. 43
4.3 Levels scale for each considered quality indicator. 44
4.4 Features extracted from attendance logs in our study. 45

5.1 Detailed information about the courses. 57
5.2 Content of the questionnaire provided to instructors. 63
5.3 Behavioral Indicators of Unknown Unknowns. 68
5.4 Model Prediction Groups. 77
5.5 Teachers' Knowledge and Interventions. 85

Chapter 1

Introduction

1.1 Motivation

Educational systems integrating artificial intelligence (AI) are transforming the landscape of education, offering personalized learning pathways, timely feedback, and data-driven insights. AI-based models have been employed in various educational applications, such as predicting student success, recommending learning materials, and providing motivational feedback [1, 2, 3, 4, 5]. While these advancements hold promising prospects, there is a growing concern about potential irresponsibility in AI-enabled educational systems. Reports of systemic biases in automated college enrollment systems and biased machine-learning evaluations for PhD applicants raise alarms about the impact of AI on educational equality [6, 7].

Efforts to address responsibility in AI applications have primarily focused on designing responsibility definitions and developing algorithms to assess and mitigate biases [8, 9]. However, responsible-aware practices should be prevalent when developing AI for education to ensure positive and equitable outcomes for learners. Therefore, understanding the challenges and needs of the educational community in integrating and monitoring responsibility in AI applications is crucial.

This thesis investigates algorithmic disparities in AI for education, with a specific focus on performance issues, including cases where AI models exhibit high confidence but make significant predictive errors. These performance problems can lead to irresponsible treatment and perpetuate inequalities in educational settings. We explore the challenges and implications of algorithmic disparities in both synchronous and asynchronous learning environments, shedding light on the significance of considering not only visible student behavior but also background knowledge, learning history, and characteristics.

1.2 Challenges

While performance analysis has been investigated in the wider AI domain, its adaptation to educational AI remains relatively underexplored. Responsibility concerns in AI for education introduce distinctive challenges, and current research in this area often consists of isolated instances. Therefore, there is a need to understand the specific challenges faced by the educational community in developing responsible AI for education and to identify the areas that require attention and improvement.

In the context of synchronous and asynchronous learning, significant opportunities exist to analyze and model student behavior. Asynchronous learning provides flexibility but lacks real-time interaction, while synchronous learning allows for immediate feedback but may require better monitoring and understanding of student engagement patterns. Analyzing and predicting student behavior in both modalities can inform teaching strategies, optimize infrastructures, and improve learning outcomes. However, the extensive data generated from these learning modes requires thoughtful analysis and interpretation to ensure its effective use.

1.3 Contributions

This thesis makes several contributions to better understand performance issues and unknown unknowns in AI for education (see also Figure 1.1):

- A comprehensive investigation of experts' challenges and needs in responsible AI for education, conducted through an anonymous survey and semi-structured interviews with educational researchers and practitioners.
- An analysis of student behavior in synchronous online learning environments to understand participation patterns, the influence of course delivery times, and implications for course attendance and quality prediction.
- A study on algorithmic disparities and unknown unknowns in student success prediction models, examining their prevalence, impact, and the feasibility of characterizing them in different instructional settings.
- An exploration of the dissonance between teacher and model predictions regarding learners at risk in a flipped course, identifying decision-making patterns, confidence levels, and intervention needs of teachers.

1.4 Outline

The following sections of this thesis will be structured as follows: in *Chapter 2* we delve into the fundamental concepts of machine learning in the context of education.

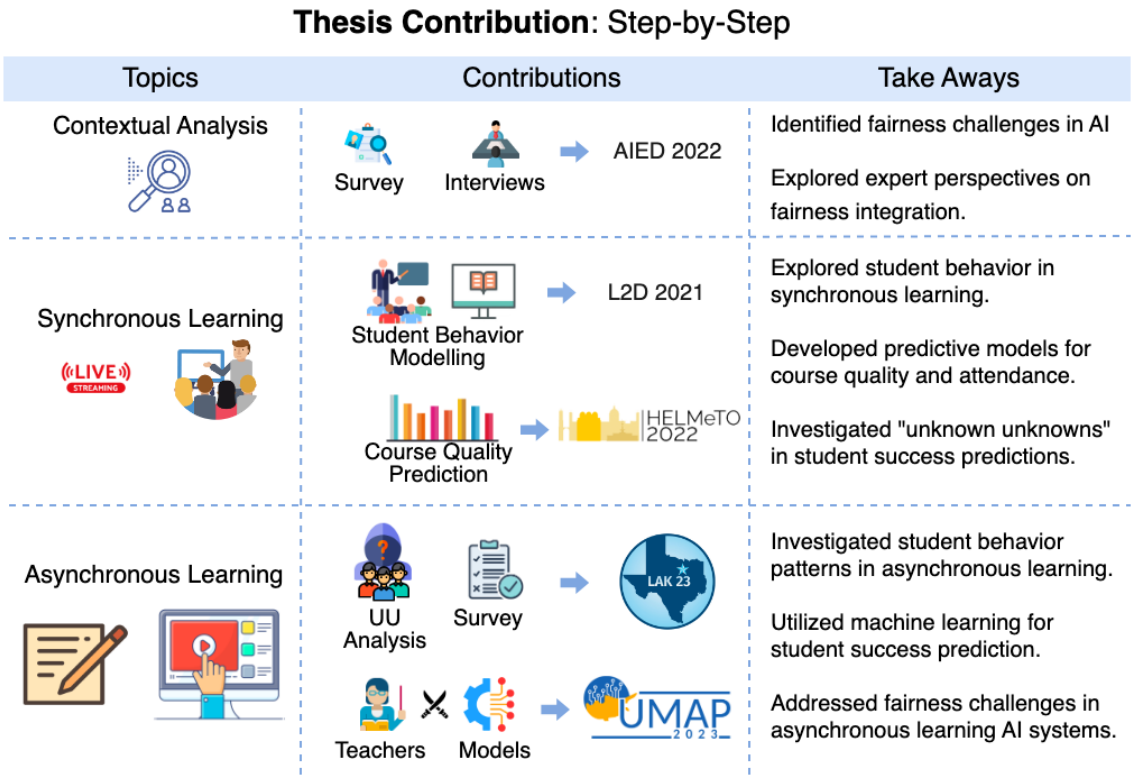


Fig. 1.1: Key Contributions of the Thesis.

It provides a comprehensive review of the methods commonly used in AI applications for education, offering essential background knowledge to understand the subsequent chapters.

Chapter 3, we present the results of our comprehensive investigation into the challenges and requirements of experts regarding fairness-aware AI in education, as well as the technical perspectives. The work presented in this chapter is a collaborative effort with Dr. Mirko Marras and Prof. Gianni Fenu from the University of Cagliari, Italy. We are proud to report that our findings have been published and presented in the prestigious Artifical Intelligence in Education (AIED 2022) Conference [10].

Chapter 4, delves into the analysis of student behavior in synchronous learning environments, exploring participation patterns and their implications. This research endeavor was conducted in collaboration with Dr. Mirko Marras, Prof. Gianni Fenu from the University of Cagliari, Italy. Our valuable findings from this chapter have been disseminated through the First International Workshop on Enabling Data-Driven Decisions from Learning on the Web (L2D 2021) [11] and the 4th International Conference on Higher Education Learning Methodologies and Technologies Online (HELMETO 2022) Conference [12].

Chapter 5, investigates algorithmic performance and unknown unknowns in student success prediction models, examining their prevalence and impact in different instructional settings. This investigation has been supported by two significant research works published in the Learning Analytics and Knowledge (LAK 2023) Conference [13] and the Conference on User Modeling, Adaptation, and Personalization (UMAP 2023) [14].

Finally, in *Chapter 6*, we summarize the main contributions of the thesis, discuss key findings, and outline future research directions.

Chapter 2

Background on Machine Learning for Education

2.1 Introduction

2.1.1 Role of ML in Education

Machine learning (ML) represents a field of computational science that empowers algorithms to acquire knowledge autonomously, without the need for constant reprogramming or external guidance. In particular, it is a technological discipline focused on developing computer algorithms capable of replicating human intelligence. Through the analysis of novel data, ML systems enhance their intelligence by recognizing and categorizing patterns and trends. This continual learning process leads to progressively improved performance. This technology has been applied in such diverse fields as pattern recognition [15], computer vision [16], spacecraft engineering [17], finance [18], entertainment [19, 20], ecology [21], computational biology [22, 23], and biomedical and medical applications [24, 25]. The most important property of these algorithms is their distinctive ability to learn the surrounding environment from input data with or without a teacher [26, 27]. The integration of machine learning within educational contexts serves as a valuable asset, aiding students, educators, and administrators in streamlining their workflows and enriching the educational experience.

Machine learning algorithms play a pivotal role in reshaping education by tailoring content, schedules, and learning objectives to individual student needs and capabilities. This level of personalization significantly enhances the efficiency and quality of both teaching and learning processes. As a result, educators can shift their focus to aspects of education that truly benefit from a human touch [28, 29, 30]. Incorporating machine learning into education empowers educators to predict future learning outcomes and adapt teaching methods accordingly. Predictive analytics harnesses the power of machine learning to identify patterns in student behavior, ultimately determining the likelihood of each student successfully completing a course or participating in extracurricular

activities [31, 32, 33]. Machine learning proves particularly invaluable at the K-12 level. It enables the early detection and prediction of behavioral issues and academic challenges with remarkable accuracy. Educators can proactively address these concerns, ensuring timely support for students before problems escalate. Additionally, ML enhances security measures and facilitates self-service tools for students and parents [34, 35]. Machine learning extends its reach to higher education, assisting institutions in predicting enrollment levels and identifying potential applicants. Moreover, ML contributes to groundbreaking research endeavors, swiftly and accurately analyzing the ever-expanding volumes of data in this domain [36, 37]. The transformative potential of machine learning isn't limited to educational institutions alone. Learning and EdTech companies harness ML to elevate learning outcomes, refine customer service, and craft targeted marketing strategies. Capabilities such as text-to-speech, translation, transcription, chatbots, and content classification are among the myriad tools enhancing the educational technology landscape [38].

2.1.2 Benefits of ML in Education

Machine learning, a facet of artificial intelligence, offers a spectrum of advantages when integrated into the realm of education. It provides a robust foundation for personalized support, enabling educators to swiftly identify and address individual students' challenges. For instance, if a teacher detects a student struggling with mathematics, machine learning steps in with insights drawn from the student's past performance, thereby assisting the teacher in fine-tuning their teaching strategies. Moreover, machine learning lends itself to the creation of tailored learning experiences. By dynamically adjusting instruction based on each student's performance [39, 40], educators can optimize their teaching methods. This not only results in time savings but also ensures an enriched learning journey for all students.

Beyond personalized support and tailored learning, machine learning contributes to enhancing overall student performance [29, 41]. Schools leverage machine learning to monitor student progress comprehensively [42]. By analyzing various data points, including test results, valuable trends come to light, empowering educators to make informed decisions about instructional improvements and timely interventions. Machine learning also serves as a beacon of efficiency and cost savings for educational institutions. It automates essential tasks like grading[43] and record management, effectively reducing operational costs. This streamlined automation of administrative processes liberates educators to dedicate more time to their primary role: teaching.

One of the most compelling advantages of machine learning is its ability to facilitate early intervention. ML algorithms excel at identifying students who might be at risk of dropping out or falling behind grade-level standards. This early detection equips educators with the insights needed to provide timely support and intervention, ultimately driving improvements in student outcomes [44, 45]. Furthermore, machine learning plays a pivotal role in understanding student behavior. Leveraging Educational Data

Mining (EDM) techniques, often powered by machine learning, educational institutions gain profound insights into students' learning behaviors and interests. This in-depth understanding forms the bedrock for designing effective teaching strategies that not only enhance performance but also curtail dropout rates [46].

In summation, these multifaceted advantages exemplify how machine learning enriches the educational landscape, delivering substantial benefits to students, educators, and institutions alike.

2.2 Educational Environments

Education occurs in diverse settings, each with its unique characteristics and requirements. Understanding these environments is essential for applying machine learning effectively in education. The following subsections provide an overview of key educational environments.

2.2.1 Traditional Face-to-Face Education

Traditional face-to-face education is the conventional form of learning that takes place in physical classrooms. In this environment, the educational process is characterized by direct, in-person interactions between educators and students. These interactions create a dynamic learning atmosphere that offers several distinct advantages [47].

One of the most noteworthy features of traditional education is the ability for students to receive immediate feedback from instructors. Whether through verbal responses to questions, discussions, or real-time assessments, this feedback loop is instrumental in reinforcing learning and clarifying doubts. Furthermore, traditional education facilitates hands-on learning experiences. In science laboratories, art studios, and vocational workshops, students can directly apply theoretical knowledge, fostering a deeper understanding of concepts and skills.

Physical classrooms encourage collaboration among students. Group projects, peer discussions, and cooperative problem-solving are inherent to this learning environment. These interactions not only enhance subject comprehension but also cultivate crucial interpersonal skills. Moreover, traditional education often follows a structured curriculum and timetable, providing a sense of routine and discipline for students. This structure can be particularly beneficial in ensuring comprehensive coverage of topics.

Beyond academics, face-to-face education supports the holistic development of students. It provides opportunities for social interaction, the formation of friendships, and participation in extracurricular activities like sports and clubs. It's essential to note that traditional education has a rich historical legacy and continues to be a predominant mode of instruction worldwide. However, advancements in technology and changing educational paradigms have led to the emergence of various alternative learning modalities,

such as online and blended learning. As a result, educational researchers often conduct studies and comparisons between face-to-face learning and these emerging models to evaluate their effectiveness, adaptability, and relevance in modern education [48, 49, 50].

This comprehensive understanding of traditional face-to-face education lays the groundwork for assessing its strengths and limitations in comparison to newer educational approaches, providing valuable insights for educators and policymakers.

2.2.2 Synchronous Learning

Synchronous learning refers to real-time, online education where students and instructors engage simultaneously. It replicates aspects of traditional classroom learning but takes place in virtual environments. Participants can interact through video conferencing, chat, and other digital tools, allowing for live discussions, group activities, and immediate feedback. One of the defining features of synchronous learning is real-time engagement. Students and instructors are connected in real-time, allowing for live discussions, questions, and responses. This real-time interaction can simulate the spontaneous exchanges often found in physical classrooms. Virtual classrooms are at the core of synchronous learning. These online spaces host a variety of interactive tools, including video conferencing, chat features, and interactive whiteboards. These tools enable instructors to present lessons, share resources, and engage students effectively [51, 52].

Live discussions are a hallmark of synchronous learning. They offer a platform for students to voice questions, share insights, and engage in debates. Instructors can facilitate these discussions, ensuring that students actively participate and contribute to the learning experience. Synchronous learning also supports collaborative group activities. Students can work together on projects, assignments, and problem-solving tasks in real time. This fosters teamwork, peer learning, and the development of critical collaborative skills [53, 54]. Like traditional classroom, synchronous learning allows for immediate feedback. Instructors can assess students' understanding during lessons and address misconceptions promptly. This feedback loop enhances comprehension and helps students stay on track [55]. Interactive learning resources are another strength of synchronous learning platforms. From multimedia presentations to live demonstrations, these resources cater to diverse learning styles and keep students engaged throughout the lesson [56].

Despite its synchronous nature, this mode of learning offers flexibility. While it has a fixed schedule, students from different geographical locations can participate, making it a viable option for distance education. Additionally, recorded sessions can accommodate students who may have scheduling conflicts. Furthermore, synchronous learning fosters a sense of community among students, despite being in virtual environments. Regular interactions and shared learning experiences create a supportive and collaborative atmosphere. In educational practice, synchronous learning can be particularly effective when

well-designed, balancing the advantages of real-time engagement with the flexibility demanded by modern learners. Educators and institutions often integrate synchronous components into blended learning models, combining the strengths of both synchronous and asynchronous approaches to create comprehensive learning experiences [57, 58].

2.2.3 Asynchronous Learning

Asynchronous learning, on the other hand, offers flexibility by allowing students to access educational content and resources at their convenience. It doesn't require simultaneous participation. Students can engage with course materials, assignments, and discussions at their own pace, making it suitable for individuals with varied schedules [59]. One of the defining features of asynchronous learning is the absence of real-time constraints. Students have the freedom to choose when and where they engage with the course materials. This flexibility is particularly beneficial for individuals with busy schedules, as it enables them to balance their education with other commitments. Online learning platforms are central to asynchronous learning. These platforms serve as repositories for educational resources, including lecture videos, readings, quizzes, and assignments. Students can access these materials 24/7, allowing them to learn at their most productive times.

Self-pacing is a key component of asynchronous learning. Students progress through the course materials independently, enabling them to spend more time on challenging concepts and move quickly through familiar topics. This autonomy over the learning process encourages self-regulation [60, 61] and time management skills [62, 63]. Discussion boards and forums are often integral to asynchronous learning environments. These platforms facilitate communication and collaboration among students and instructors. While not in real-time, these discussions provide opportunities for students to ask questions, seek clarification, and engage in meaningful conversations about course content.

One significant advantage of asynchronous learning is its accessibility [64]. Students from diverse geographical locations can participate, overcoming the barriers of time zones and physical distances. This inclusivity enhances the diversity of perspectives in the learning community. Asynchronous learning also accommodates various learning styles. Students can choose the format that best suits their preferences, whether it's reading text-based materials, watching video lectures, or participating in interactive simulations. This flexibility caters to a broad range of learners. While instructors are not present in real-time during asynchronous learning, they remain actively involved. They design and curate course materials, provide clear instructions, and set deadlines for assignments and assessments. Instructors also participate in discussion boards, answer questions, and offer guidance. Assessment in asynchronous learning is often based on assignments, quizzes, and exams that students complete within specified time-frames. This approach allows instructors to gauge students' understanding of the material and provide feedback for improvement [65].

In summary, asynchronous learning offers a flexible and inclusive educational experience. It caters to students' diverse needs and schedules, promotes independent learning, and fosters a sense of self-regulation. When thoughtfully designed, asynchronous courses provide valuable educational opportunities for a wide range of learners, contributing to the ever-evolving landscape of online education.

2.3 Data Mining Techniques

Data mining techniques play a pivotal role in educational data analysis, particularly in the classification and prediction of students' outcomes. These methods are employed to distribute datasets into distinct classes, facilitating predictions about future data based on predefined categories. In this section, we'll provide an overview of key classification techniques, used in previous research (as shown in Table 2.1), for predicting students' performance.

One of these techniques is Decision Trees. Decision Trees are graphical models that use a tree-like structure to make decisions. In education, these trees are used to classify students into different categories based on features such as grades, attendance, or demographic information. Decision Trees are interpretable and can help educators identify important factors that affect student performance. Another technique is K-Nearest Neighbors (K-NN). K-NN is a classification algorithm that assigns a class label to a data point based on the majority class among its k-nearest neighbors. In education, it can be used to predict a student's performance by looking at the performance of their closest peers in terms of academic history, interests, or study habits. Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. In educational contexts, it can be used to make more robust predictions about student outcomes by considering various factors and reducing the impact of noise in the data.

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. In education, SVM can be applied to predict student outcomes by finding a hyperplane that best separates different classes of students based on features such as test scores, attendance, or study time. Artificial Neural Network

Acronym	Description
Decision Trees	Graphical models for classifying students based on features like grades and more.
K-NN	Assigns class labels based on peers' performance in academic history, interests, etc.
Random Forest	Ensemble method for robust student outcome predictions by considering various factors.
SVM	Separates students into classes using a hyperplane based on features like test scores.
ANN	Complex pattern recognition using interconnected nodes for predicting student performance.
Naïve Bayes	Probabilistic classification for predicting student outcomes based on attributes.

Table 2.1: Data Mining Techniques for Student Performance Prediction.

(ANN), inspired by the structure and function of the human brain, consists of layers of interconnected nodes (neurons) and is used for complex pattern recognition tasks. In education, ANNs can be employed for predicting student performance by processing a vast amount of data and identifying intricate patterns and trends. Naïve Bayes is a probabilistic classification algorithm based on Bayes' theorem. It assumes that features are conditionally independent, simplifying the calculation of probabilities. In educational applications, Naïve Bayes can be used to predict student outcomes by estimating the likelihood of a student belonging to a particular class based on their attributes and behaviors.

These are some of the fundamental data mining techniques utilized in educational data analysis. Each technique has its strengths and limitations, making them suitable for different types of educational predictive tasks. The choice of technique often depends on the specific problem and the nature of the available data.

2.4 Data Mining Tools

In the field of Educational Data Mining (EDM), several tools and software platforms have emerged to facilitate data analysis, predictive modeling, and decision-making processes. These tools offer educators and researchers powerful capabilities for extracting valuable insights from educational datasets. Below are some of the commonly used tools:

R is a popular open-source programming language and environment for statistical computing and graphics. It offers a wide range of packages and libraries specifically designed for data mining and machine learning in education. Researchers and educators often use R for data preprocessing, visualization, and building predictive models[1].

Python is another versatile programming language widely adopted in EDM. Its extensive ecosystem of libraries, including scikit-learn, pandas, and TensorFlow, makes it suitable for tasks such as data analysis, machine learning, and natural language processing in the educational domain[2].

Weka is user-friendly data mining software that provides a graphical user interface for building and evaluating machine learning models. It offers a wide variety of data preprocessing, classification, clustering, and visualization tools, making it accessible to educators and researchers without extensive programming experience[3].

RapidMiner is an integrated data science platform that simplifies the entire data mining workflow, from data preparation to model deployment. It offers educational institutions

[1] https://www.r-project.org/

[2] https://www.python.org/

[3] https://www.cs.waikato.ac.nz/ml/weka/

and researchers an intuitive interface for building and deploying predictive models[4].

KNIME is an open-source platform for data analytics, reporting, and integration. It allows users to create data pipelines and execute data mining workflows through a visual interface. KNIME's flexibility and extensibility make it a valuable tool for educators and researchers in EDM[5].

Orange is a user-friendly open-source data visualization and analysis tool. It provides a visual programming interface for constructing workflows and analyzing data, making it suitable for both beginners and experts in EDM[6].

IBM SPSS is a widely used statistical software package that includes advanced analytics capabilities. In the context of education, it can be employed for data analysis, predictive modeling, and reporting[7].

Tableau is a popular data visualization tool that helps educators and researchers transform complex educational data into interactive and easy-to-understand visualizations. It enables users to gain insights from data quickly and effectively[8].

These tools offer diverse capabilities for data mining, predictive modeling, and data visualization in educational settings. The choice of tool often depends on the specific requirements of the educational data analysis project and the expertise of the users involved. For this thesis, Python served as the primary tool for data analysis and modeling.

2.5 Data Mining Process

In this section, we delve into the methodologies employed to bridge the gap between raw data and practical applications within the scope of AI in education. The methods outlined here encompass the transformative journey from data collection to the development of AI-enabled educational systems, reflecting the essence of translating research insights into real-world impact. The steps detailed below in the data mining process, are also presented in Figure 2.1

Step 1: Data Collection. The foundation of any data-driven research in AI for education begins with data collection. This process involves acquiring diverse datasets encompassing student demographics, academic performance, engagement metrics, and other relevant attributes. Data preprocessing is a crucial initial step, entailing tasks like data cleaning, normalization, and feature extraction. It ensures that the data is in a suitable format for subsequent analysis.

[4] https://rapidminer.com/

[5] https://www.knime.com/

[6] https://orangedatamining.com/

[7] https://www.ibm.com/it-it/spss

[8] https://www.tableau.com/

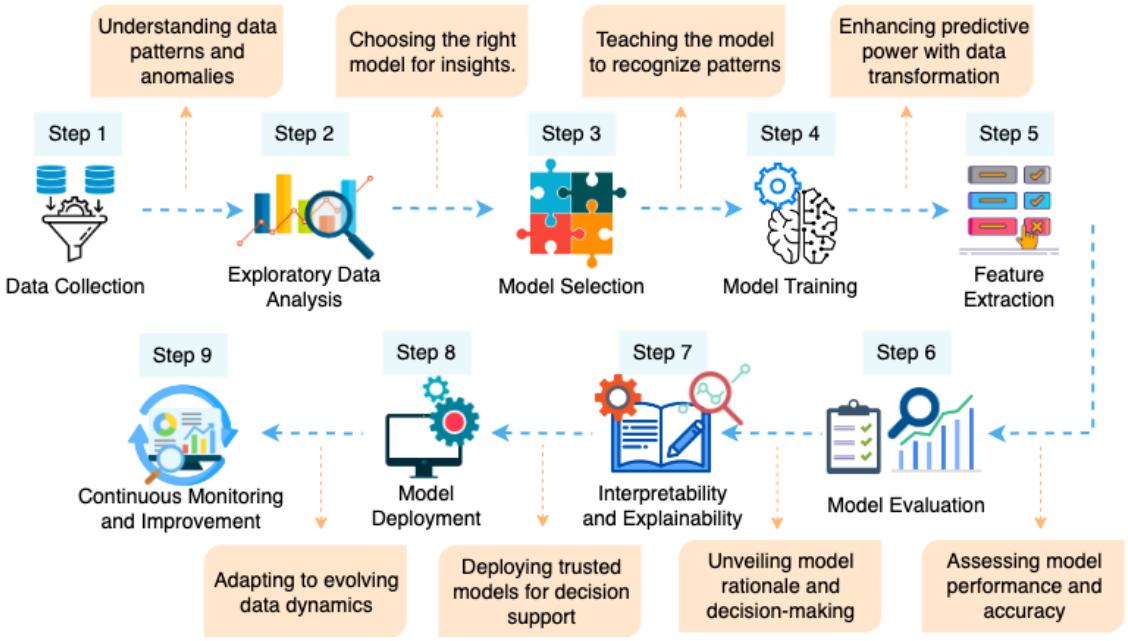


Fig. 2.1: Interconnected steps in the data mining process for AI-driven educational systems

Step 2: Exploratory Data Analysis. Exploratory Data Analysis is an indispensable phase aimed at understanding the inherent patterns, distributions, and anomalies within the collected data. Statistical and visualization techniques are leveraged to unearth valuable insights, identify trends, and highlight potential correlations among variables. This stage contributes significantly to guiding the subsequent analysis.

Step 3: Model Selection. The selection of appropriate machine learning models hinges on the research objectives and the nature of the data. Commonly employed algorithms include decision trees, support vector machines, random forests, and neural networks, among others. The chosen models are tailored to address specific tasks, such as predicting student outcomes or identifying engagement patterns.

Step 4: Model Training. The selected machine learning models are trained on a portion of the dataset, utilizing techniques like cross-validation to ensure robustness and generalizability. During this phase, models learn to recognize underlying patterns and relationships within the data.

Step 5: Feature Extraction. Feature extraction involves the creation of new variables or the transformation of existing ones to enhance the predictive power of machine learning models. This process can uncover latent insights and improve the accuracy of predictions, making it an essential component of the methodology.

Step 6: Model Evaluation. Rigorous evaluation of model performance is conducted using metrics tailored to the specific research goals. Common evaluation metrics include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Models are assessed for their ability to make accurate predictions and provide actionable insights.

Step 7: Interpretability and Explainability. In the context of AI in education, interpretability and explainability of models are paramount. Interpretability tools, such as SHAP (SHapley Additive exPlanations) values, LIME (Local Interpretable Model-Agnostic Explanations), and feature importance analysis, are employed to elucidate the rationale behind model predictions. This ensures that stakeholders can trust and comprehend the AI-driven decisions.

Step 8: Model Deployment. Successful machine learning models are deployed within educational systems, where they facilitate data-driven decision-making. Deployment involves integrating models into existing educational technologies or platforms, ensuring seamless interaction with educators, administrators, and students.

Step 9: Continuous Monitoring and Improvement. The journey from data to applications does not culminate with deployment; it evolves into a cycle of continuous monitoring and improvement. Models are subject to ongoing evaluation, recalibration, and fine-tuning to adapt to evolving educational landscapes and data dynamics.

Throughout these methodological phases, ethical considerations remain a paramount concern. The research adheres to ethical guidelines, promoting fairness, transparency, and accountability in AI-enabled educational systems.

In essence, this section outlines the systematic progression from data acquisition and preprocessing to the development, deployment, and refinement of machine learning models within the educational context. These methods form the critical bridge that translates data insights into actionable solutions, ultimately contributing to the advancement of AI in education.

2.6 Educational Applications

In this section, we explore diverse applications of AI in education, showcasing how machine learning and data-driven approaches are harnessed to address critical challenges and enhance various facets of the educational landscape. These applications underscore the multifaceted impact of AI technologies on students, educators, and institutions alike.

One of the foremost applications of AI in education is predicting student success. Machine learning models analyze historical student data, including academic performance, engagement metrics, and demographics, to forecast outcomes. Early identification of

at-risk students allows educators to intervene proactively and provide tailored support, ultimately improving retention rates and student success [66, 67]. AI-powered personalized learning systems adapt content, pace, and instructional strategies to individual student needs. These systems leverage data analytics to understand students' strengths and weaknesses, ensuring that learning experiences are tailored for optimal comprehension and engagement [68, 69]. Student modeling involves creating detailed profiles of individual students based on their interactions with educational platforms. These models track learning progress, identify knowledge gaps, and provide recommendations for further study. This application enhances both teaching and learning by offering real-time insights into student performance [70, 71].

AI algorithms analyze student behavior and preferences to recommend relevant educational content. Whether suggesting reading materials, exercises, or supplementary resources, content recommendation systems foster self-directed learning and engagement [72]. Machine learning streamlines the grading process by automating the evaluation of assignments, quizzes, and exams. This not only reduces the workload on educators but also ensures consistency and objectivity in grading practices [73]. AI-driven data analysis informs curriculum development and enhancement. Educators and institutions can use insights from machine learning to optimize course content, delivery methods, and assessment strategies, resulting in more effective teaching and learning [74].

Learning analytics harness AI to scrutinize vast educational datasets. This application extracts actionable insights, such as identifying effective teaching strategies, optimizing resources, and enhancing overall learning experiences [75]. Gamification incorporates game elements into educational contexts to boost engagement and motivation. AI algorithms can tailor gamified experiences to individual student preferences and learning objectives [34]. AI streamlines administrative tasks in educational institutions, from student enrollment and resource allocation to facility management and budget planning. Automation of these processes optimizes efficiency and resource utilization. AI-driven language processing tools facilitate language acquisition by offering real-time translation, pronunciation feedback, and grammar correction. These applications are particularly valuable in language learning courses [76].

AI-driven assistive technologies support students with disabilities, providing tools such as speech recognition, text-to-speech, and screen readers to ensure equal access to educational content. Institutions leverage predictive analytics to forecast enrollment trends, allocate resources effectively, and plan for future academic offerings. AI-powered chatbots offer immediate support to students and educators. They can answer questions, provide guidance, and offer feedback, enhancing the overall educational experience [77].

These diverse applications demonstrate the transformative potential of AI in education, from improving student outcomes and engagement to enhancing administrative efficiency. By harnessing the power of machine learning and data analytics, educational institutions are poised to revolutionize teaching and learning in the digital age.

Chapter 3

Needs and Challenges in Machine Learning for Education

This chapter presents the findings of our investigation into experts' challenges and needs in performance analysis AI for education.

3.1 Introduction

Artificial intelligence (AI) integrated into educational systems is significantly impacting the quality of education. Examples of AI-based models integrated so far include early predictors of student success [1], clustering techniques for learner modeling [2], intelligent tutoring and scaffolding [3], agents for motivational diagnosis and feedback [4], and models for recommending peers or learning material [5]. However, alongside this growth, there is a growing concern about the potential of AI to exacerbate unfairness in educational applications. Mainstream media has reported systemic unfair behaviors of some AI-enabled educational systems, such as automated college enrollment systems that exhibit biases based on ethnicity, gender, or age [6], or machine-learning systems for evaluating PhD applicants that perpetuate existing inequalities in the field [7].

Efforts to address fairness in educational applications have mainly focused on designing fairness definitions [8] and developing algorithmic methods to assess and mitigate biases based on these definitions [9]. Some studies have also examined fairness in educational AI systems from social and psychological perspectives [78]. While there are already some research studies on fair AI, they often represent isolated examples. For the resulting AI applications to have a positive impact on education, fairness-aware practices should become common when developing educational applications that leverage AI. Understanding the actual challenges and needs for developing fairer AI for education is therefore crucial.

Creating AI-based educational systems presents unique challenges that are not commonly encountered in other domains of AI [79]. Although fairness has received attention

in the broader AI field [80, 81, 82, 83], only a few studies have specifically investigated challenges and needs for creating fairer AI by directly consulting experts [84, 85]. Unlike previous studies that focused on public-sector and commercial AI practitioners, our study focuses on educational researchers and practitioners who are incorporating AI into their work but are relatively new to considering fairness. Integrating fairness considerations, beliefs, practices, motivations, and priorities may be less clear in these educational contexts and cultures.

In this chapter, we are interested in investigating the challenges and needs faced by the educational community, whose products have a direct impact on individuals' education, in integrating and monitoring for unfairness and taking appropriate action. Through an anonymous survey of 136 educational researchers and practitioners who have published their research in top-tier educational conferences in 2021, we analyze their existing opinions, experiences, challenges, and needs regarding the development of fair educational AI. Additionally, we conduct semi-structured interviews with 29 of these experts to delve deeper into the key themes identified in the survey. To our knowledge, this is the first systematic investigation of experts' challenges and needs related to fairness in educational AI.

Through our investigation, we identify a range of real-world needs that have not been extensively addressed in the literature thus far, as well as several common areas. For instance, unlike the broader AI field, large-scale data collection is not always considered a solution in educational AI due to complex biases driven by local contextual factors. Research teams also struggle with identifying sub-populations and forms of unfairness that need to be considered for specific applications, indicating their own blind spots. Moreover, while fair educational AI has predominantly focused on data collection issues, assessment and debiasing of unfairness are equally crucial, necessitating continuous fairness assessment at all stages of the development pipeline. Given the context and application-dependent nature of fairness, there is an urgent need for domain-specific educational resources, metrics, processes, and tools, including open data and source code for public scrutiny and participatory processes for fairness checking. Another area that requires attention is the development of auditing processes and tools to bring fairness issues to light. Based on our findings, we highlight opportunities to have a greater impact on the landscape of fair educational AI. Given the complexities associated with these challenges and the crucial need for equitable development of educational AI, we present the following research questions:

1. **RQ1:** What are the significant challenges faced by individuals and organizations interested in the practical application of these techniques within real-world settings, particularly in the realm of education and AI?

2. **RQ2:** Considering the multifaceted challenges identified in the first question, what are the current, pressing needs and requirements for effectively addressing them in the pursuit of advancing responsible AI within educational contexts?

3.2 Methodology

In our study, we took several steps to ensure a comprehensive understanding of the challenges and needs for addressing fairness in the development of educational AI.

3.2.1 Survey Study Implementation

Initially, we conducted an anonymous online survey to gather a broad sense of these issues. To ensure a representative sample, we employed a systematic recruitment process rather than relying on an arbitrarily selected population.

We manually scanned the proceedings of top-tier educational conferences held in 2021, including AIED, EAAI, EC-TEL, EDM, ICALT, ITS, LAK, and L@S. Additionally, we considered authors of papers in special issues about fair educational AI in IJAIED. This rigorous approach allowed us to identify authors who had papers accepted at these conferences and directly emailed them the survey between September and December 2021. Furthermore, we encouraged them to share the survey with their colleagues working on educational AI within their organizations. In total, we contacted 2,175 experts, and a noteworthy 136 individuals (6%) completed at least one section beyond the demographic questions. A description of the respondents is provided in Fig. 3.1a.

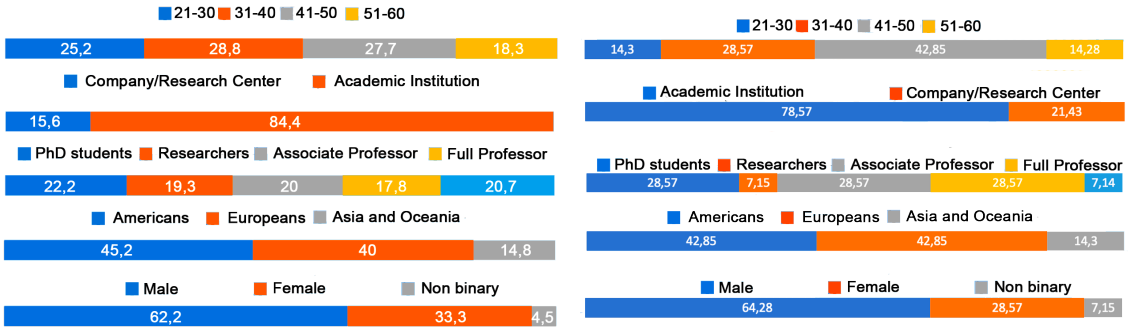
The survey was structured as a Google Form, and we designed specific questions to explore the prevalence and generality of emerging themes. Initially, we gathered demographic information to gain insights into the respondents' backgrounds, including their technological areas and roles. Subsequently, the survey consisted of sections branching into various stages of the educational AI development pipeline[1]. In each section, participants were asked about their opinions, challenges, and support requirements regarding fairness. Open-ended response options allowed respondents to elaborate on their arguments. Finally, we requested their email addresses in case they were willing to participate in subsequent interviews.

By employing this comprehensive survey methodology, we aimed to collect diverse perspectives and insights from the educational AI community.

3.2.2 Interview Study Implementation

In order to validate and deepen our findings from the previous survey, we conducted a series of semi-structured, one-to-one interviews with selected participants. To recruit interviewees, we reached out to the experts who had completed the survey and expressed their willingness to participate in follow-up interviews by providing their email addresses. Out of the 136 survey respondents, 29 individuals (21%) agreed to take part in this second step, representing a diverse range of research teams. We made an effort

[1] A pdf copy of the survey questions is available at `https://bit.ly/FairAIEdSurvey`.



(a) Survey **(b)** Interview

Fig. 3.1: Sample population statistics for our survey and interview process.

to include individuals in different roles within the same team to capture diverse perspectives. The interviews were conducted remotely, as the participants were located in various countries.

During each interview, we reminded the participants of the purpose of our research and then focused on their awareness of the ongoing debate and research on demographic fairness in educational AI. We also explored the most important challenges and open questions in the field, as indicated in questions 7 and 8 of the survey. Furthermore, we asked participants to provide more details about the educational AI applications they were working on and the target users of these applications, as mentioned in question 8 of the survey. We delved into the participants' perspectives on fairness and whether it is regularly considered in their workflow. We aimed to understand their interpretation of fairness within their specific application context, as mentioned in question 9 of the survey. The discussion during the interviews often referred back to the participants' survey responses.

To further explore fairness in the educational AI development pipeline, we presented survey questions related to each stage, from data collection to dataset design, and the assessment and potential mitigation of fairness issues (questions 10 to 13 of the survey). For each stage, we posed a broad opening question aligned with the corresponding survey question, followed by specific follow-up questions based on the participants' survey responses. This approach encouraged participants to provide more in-depth insights into their practices and reflections. Through this series of interviews, we aimed to validate and enrich the insights gained from the survey. By gaining a deeper understanding of the perspectives and experiences of the educational AI community, we sought to strengthen the findings and ensure a comprehensive understanding of the challenges and needs related to fairness in educational AI.

3.2.3 Survey and Interview Data Analysis

After collecting data from both the survey and interview phases, we proceeded with the analysis of the responses. Each survey and interview entry was assigned a unique ID, allowing us to link corresponding survey and interview data from the same participant.

To analyze and synthesize the findings, we followed a standard methodology used in contextual design. We conducted interpretation sessions and employed affinity diagramming, a technique commonly used in qualitative research (e.g., see also [86]). Our approach involved a bottom-up process of generating codes based on individual text segments, followed by grouping these codes into higher-level themes. Importantly, the themes that emerged from the analysis were derived directly from the data rather than being imposed on the responses.

Through this iterative process of affinity diagramming, we identified key themes that encapsulated the participants' perspectives and experiences. These themes, which we will present in the following section, provide valuable insights into the challenges and needs surrounding fairness in educational AI.

3.3 Results and Discussion

In the following discussion, we will address the current challenges and needs related to fairness in educational AI. These insights are organized into top-level themes, which are based on both the survey questions and the in-depth interviews. The interviews served to confirm and enrich the responses obtained from the survey, providing a more comprehensive understanding of the subject. The resulting affinity diagram helped us categorize the challenges and needs into specific sub-themes.

The themes discussed cover various aspects, starting with the challenges and needs surrounding data collection and modeling, as highlighted in questions 9—10. We explore the issues related to detecting and mitigating unfairness in educational AI, as presented in question 11. Additionally, we examine the provision of fairness guarantees, as outlined in question 12, and the importance of holistic fairness auditing, as discussed in question 13. These themes are accompanied by systemic aspects such as team composition, cross-organizational collaborations, and the maturity of educational AI, as addressed in question 7.

Within each top-level theme, we further present selected sub-themes that shed light on specific areas of concern. It is important to note that our study primarily aims to identify open questions and highlight areas that require further exploration. Providing comprehensive answers to these challenges necessitates ongoing discussions and collaborative efforts within the research community as a whole. Our study serves as a catalyst for these discussions and encourages further work in this important field.

3.3.1 Challenges and needs in data collection and grouping

Cultural dependencies in demographic representation. Many participants acknowledged that researchers themselves, who are often not demographically representative of their societies, tend to involve individuals within their immediate circles, such as their own students, who may also lack diversity. In question 10 of the survey, a participant highlighted the challenge of collecting data that adequately represents different contexts, including countries, universities, and society, due to variations in culture, viewpoints, and rules. During the interview, the same participant further commented on how research tends to focus on specific countries and is predominantly conducted in English, resulting in findings that are more representative of certain societies and educational systems. Overall, it was commonly emphasized that no dataset fully encompasses the diversity of the population, inevitably leaving some individuals underrepresented. This cultural dependency poses a significant challenge when striving for fairness in educational AI, as it requires considering and addressing the inherent biases and limitations present in the data collection process.

Biases driven by reasons to be understood in the local context: In contrast to the broader field of AI, several participants expressed the belief that large-scale data collection may not effectively capture the nuances of fairness in educational AI. They emphasized that biases in educational AI are influenced by complex and context-specific factors. As stated by a participant in question 10 of the survey, *"biases in education are driven by complex reasons to be understood locally."* This highlights the need for localized data collection efforts and the importance of sharing data practices to address fairness issues effectively.

Hidden relationships between demographics and learning variables: Participants encountered difficulties in identifying the underlying issues that contribute to fairness challenges. During an interview, a participant mentioned that in some cases, ethnicity itself may not directly cause differences in how students interact with educational software or the resulting data. Instead, it is the students' life experiences, correlated with ethnicity, such as facing discrimination, that influence their engagement. It was noted that different demographic groups might respond differently to psychological measures and educational interventions. In light of this, a participant suggested during the interview that educational AI models might need to be demographically stratified. Overall, challenges were identified in understanding what demographic attributes truly represent and how experts can effectively measure and address their impact on fairness in educational AI.

Giving individuals continuous control of their data: A significant number of participants emphasized the importance of individuals having complete access and control over their data, including any new data generated about them. It was widely recognized that individuals should be able to manage their data in a way that ensures confidential-

ity and prevents unintended sharing. As one participant suggested, "access should be controlled in such a way that confidential information will not be inadvertently shared beyond their control," building upon their response in the survey. The affinity diagram revealed a clear need for supporting tools that inform users about the data being used by the system and for what purposes. A participant envisioned in the survey that these tools could allow users to selectively enable or disable the use of specific data by the system. Challenges and needs were identified regarding the development of mechanisms that empower individuals to exercise control over their data within educational systems.

Our study suggests the importance of localized data collection efforts. We emphasize the need to capture the nuances of fairness issues by considering local contexts, cultures, and viewpoints. Additionally, our findings underline the significance of involving diverse and representative groups in the data collection process to ensure a more balanced and culturally sensitive dataset.

3.3.2 Challenges and needs of fairness-aware technical pipelines

Continuous fairness assessment at all stages of the pipeline: Participants consistently emphasized the importance of integrating fairness considerations throughout the entire development pipeline of educational AI systems. They stressed that fairness should be a fundamental aspect from the outset, influencing choices related to data collection, optimization criteria, and interventions. One participant highlighted the need for clearly defined aims and objectives of data collection, emphasizing the importance of explicit discussions and negotiation with participants. The inclusion of fairness should be seen as an integral part of the design process, involving expertise in fairness and ensuring its integration at every stage. Protocols and guidelines are needed to facilitate the incorporation of fairness considerations throughout the pipeline.

Understanding and acknowledging weaknesses of the system: Participants recognized the significance of understanding the strengths and weaknesses of educational AI systems. It was noted that achieving full transparency or explainability may be challenging, but efforts should be made to comprehend the scope and limitations of the underlying data. Participants suggested informing users about the limitations of the system and its accuracy variations among different demographic groups. By acknowledging these aspects, a better understanding can be gained regarding the capabilities and limitations of the systems.

Reducing frictions between model effectiveness and fairness: Balancing prediction accuracy with fairness was identified as a major challenge. Participants expressed concerns that using demographic features directly for accurate predictions may not align with fairness goals, as those features may encode biases. Alternative practices were suggested that achieve comparable performance without relying heavily on demographic features. Additionally, debates arose regarding whether the benefits of a model that per-

forms well for one group should be withheld from another. Exploring approaches that promote fair usage while maintaining high model performance is crucial.

Creating cross-institutional frameworks for addressing fairness: The need for cross-institutional collaboration and the establishment of unified frameworks for data collection and fairness-aware model evaluation was highlighted. Participants proposed the formation of consortia consisting of organizations from different countries, such as universities and companies, to develop a shared framework. Trust-building among government, institutions, researchers, and practitioners was deemed essential for accessing sensitive data, while ensuring compliance with privacy regulations and using de-identified data in educational systems. However, participants acknowledged the challenges associated with leveraging data, even when anonymized, to improve educational systems.

Our study underscores the need for continuous fairness assessment at all stages of the pipeline. We advocate for the integration of fairness considerations from the outset, influencing choices related to data collection, optimization criteria, and interventions. This approach ensures that the individual strengths and limitations of students are considered throughout the development process.

3.3.3 Challenges and needs in providing fairness guarantees

Opening data and source code for public scrutiny: Participants expressed the importance of transparency and public scrutiny in the development of educational AI systems. They emphasized the need for developers to publish or release models, analyses, and related resources for public examination, particularly when concerns about fairness arise. Sharing data, source code, and pre-trained models in open online repositories was seen as an essential practice. However, guidelines and directives regulating this sharing process are necessary to navigate the tension with copyright and intellectual property rights.

Fairness should not be a property of the model only: Participants emphasized that fairness should not be limited to the underlying predictive model but should extend to the overall service provided to users within the educational ecosystem. They highlighted the need for guidelines and practices that embed fairness as a constraint or metric for the underlying model and as a key indicator for the service itself. Fairness should be considered throughout the entire system's deployment, not just in the design of the model.

Showing explicit evidence of the system's potential unfair impacts: Participants stressed the importance of institutions adopting educational AI systems having access to evidence that supports the claim of fairness. There were differing opinions regarding the extent to which transparency should be provided to students. Some participants sug-

gested that students should not be made explicitly aware of demographic considerations to avoid invoking stereotype threat. However, others argued that students have the right to know how the system works and be informed about any fairness shortcomings. The level of transparency, accountability, and explainability should be proportional to the impact of the system, with more significant decisions requiring greater accountability.

Creating participatory processes for fairness checking: Participants proposed the establishment of independent third-party entities or ethical commissions responsible for assessing the fairness of educational AI systems. These entities would provide sample data to the system and evaluate its fairness. Additionally, the involvement of learner advocates or conducting exploratory research to uncover potential detrimental effects on learner success was suggested. The exact mechanisms and stakeholders involved in these participatory processes are open questions to be addressed by the research community.

Regulations for defining responsibilities around fairness issues: Participants emphasized the importance of defining clear responsibilities and consequences for ensuring fairness in educational AI systems. They drew parallels to service level agreements in cloud services, where failure to meet guarantees may result in financial penalties. It was suggested that guarantees of fair treatment and certification of fairness according to specific variables should become mandatory in fair educational AI systems. Regulations and guidelines are needed to establish accountability and consequences for not upholding fairness guarantees.

Our research findings recommend the formation of consortia consisting of organizations from different countries. These consortia would develop a shared framework, emphasizing trust-building among government, institutions, researchers, and practitioners. Such frameworks are essential for accessing sensitive data while ensuring compliance with privacy regulations and fostering transparent and equitable educational AI outcomes.

3.3.4 Challenges and needs of a more holistic fairness auditing

Human-centered evaluation of fairness: Participants highlighted the importance of human-centered evaluation of fairness, which should involve multiple levels of analysis. This evaluation process would incorporate statistical metrics, expert audits of system design and training data sets, and meetings with stakeholders representing the most impacted groups. The goal is to ensure that the evaluation protocol is tailored to the specific educational context. While automation can play a role in parts of the evaluation process, stakeholders should remain actively involved.

Creation of tools that allow stakeholders to audit models: Transparency of educational AI systems to end-users is crucial, and participants expressed the need for stakeholders to be able to analyze system data for fairness and outcomes. Some participants

noted that this auditing process would require experts, as it may be resource-intensive for other stakeholders. It is important for students (or instructors) to understand why they receive certain predictions from the system so that they can reflect and respond constructively. However, there may be challenges in explaining AI-driven predictions to students, especially younger ones. Efforts should be made to ensure that students have a high-level understanding of what the system does, potentially through specific training.

Contextualized and application-specific properties to inspect: Participants expressed doubts about the generalizability of fairness metrics and protocols from the broader AI community to educational AI systems. They emphasized the need to investigate and develop a fairness spectrum specifically tailored to the educational field. Context-specific frameworks, adapted to different educational contexts and applications, should be developed, taking into account local data privacy and protection laws. The unique characteristics of the educational field, which is highly human-centered, require frameworks that are aligned with its specificity rather than relying on black-box approaches from other domains.

Long-term learning-related evaluation of fairness: Participants argued against an overly computational definition of fairness that focuses on demographic differentials instead of recognizing and addressing the strengths and weaknesses of individual students to help them reach their full potential. They questioned the overemphasis on metrics and protocols, noting that unfair outcomes could be deemed fair solely based on model performance on these metrics. Instead, participants proposed evaluating systems based on broader educational goals, looking beyond immediate intended effects (e.g., whether an auto-generated hint helps students answer a question) and considering long-term outcomes in students' overall educational performance.

3.3.5 Challenges and needs in team blind spots and practices

Support in the selection of demographic groups to consider: Participants expressed the challenge of determining which demographic attributes to consider in an analysis. There were differing opinions on whether certain demographic attributes, such as gender, are relevant or irrelevant for certain problems. The lack of consideration for socioeconomic characteristics in many studies was also highlighted as a potential bias in resulting models. Participants noted that datasets often lack fair observations and may still contain biases due to human decision-making. Establishing standards for the demographic groups to consider is necessary to address these challenges.

Building social and multi-disciplinary awareness in teams: Participants emphasized the need for teams developing educational AI models to have social science training and an understanding of the socio-cultural implications of their algorithm designs. It was noted that technical individuals without such training may prioritize computational

efficiency over social justice considerations. Participants called for equity training and awareness among developers and researchers. Inclusive and diverse teams were deemed necessary to incorporate multiple perspectives, including those from social sciences, to understand the validity and implications of collected variables in AI models that impact people.

3.4 Findings and Recommendations

In this chapter, we present the findings of our systematic investigation, which aimed to understand the challenges and needs faced by expert teams in developing fairer educational AI systems. Our research sheds light on the technical and organizational barriers that experts encounter, despite their motivation to improve fairness in educational applications. Though researchers and practitioners are already grappling with biases and unfairness in educational AI systems, research on this topic is rarely guided by a common understanding and view of the faced challenges and needs. In this work, we conducted the first systematic investigation of experts teams' challenges and needs for support in developing fairer educational AI. Even when experts are motivated to improve fairness in their educational applications, they often face technical and organizational barriers. We highlight a few emerged aspects below.

Findings RQ1. *Challenges in applying AI in education include complex data collection, driven by cultural and local factors. Understanding how demographics relate to learning variables is challenging, as is ensuring data control and empowerment tools for individuals. Fairness issues extend beyond AI models to the entire user service, requiring explicit evidence and clear responsibilities.*

Future research should also support experts in collecting and curating high-quality datasets, with an eye towards fairness in downstream AI models, reducing cultural dependencies in demographic representation. Moreover, large-scale data collection should be paired with an in-depth description of the local contexts, since biases are driven by complex reasons to be understood locally. Localized and causal data collection paired with data sharing practices are needed, posing attention in giving individuals control of their data. Though fair educational AI has mainly focused on data collection, assessment and debiasing of unfairness is also an important area of work. Challenges and needs in this area include having continuous fairness assessment at all stages of the pipeline, understanding and acknowledging the potential weaknesses of the system, reducing frictions between model effectiveness and fairness, and creating cross-institutional frameworks for addressing fairness.

Findings RQ2. *Addressing AI challenges in education necessitates transparency, data sharing, and fairness integration in the entire service. Evidence of fairness and consideration of long-term learning outcomes are vital. Third-party fairness assessment and clear accountability are essential. Teams need help in selecting demographic factors and building multi-disciplinary awareness.*

Domain-specific educational resources, metrics, processes, and tools are urgently needed. Challenges and needs in this perspective include, among others, practices for opening data and source code for public scrutiny, including fairness not only as a property of the AI model, showing explicit evidence the system's potential unfair impacts, creating participatory processes for fairness checking, and defining responsibilities around fairness issues. The development of processes and tools for fairness-focused auditing is also important, to surface fairness issues in complex, multi-component educational AI systems. Among others, challenges and needs include fostering a more focused human-centered evaluation of fairness, contextualized and application-specific tools for auditing, and long-term learning-related auditing of fairness. Finally, another area with several challenges and needs concern the teams working on educational AI. Among others, supporting the team in the selection of the demographic groups to consider and building multi-disciplinary awareness in teams are two of the more relevant aspects to work on.

The rapidly growing area of fairness in educational AI presents many challenges and needs. The resulting systems are increasingly widespread, with proved potential to amplify social inequities, or even to create new ones. As research in this area progresses, it is urgent that research agendas are aligned with the challenges and needs of those who affect and are affected by educational AI systems. We view the directions outlined in this work as critical opportunities for the AI and the educational research communities to play more active, collaborative roles in making real-world educational AI systems fair.

Chapter 4

Analyses of Algorithmic Performance in Synchronous Learning

In Chapter 3, we ventured into the intricacies of AI techniques for data collection and model fairness in educational AI systems. Now, in this Chapter, our focus narrows to synchronous learning environments. We delve into algorithmic performance within this context, specifically into student behavior modeling, course attendance prediction, and course quality prediction in AI-enabled educational systems. This chapter is dedicated to dissecting the challenges inherent to algorithmic performance and presenting our research findings pertaining to data analysis and predictions within the synchronous learning domain.

4.1 Error Analysis on Student Behavior Modelling

4.1.1 Introduction

The Covid-19 pandemic has compelled educational institutions to transition from traditional face-to-face classroom learning to online environments. Online learning refers to educational approaches that utilize multimedia and Internet technologies for teaching and learning purposes. Initially, this strategy was adopted to support students who were unable to physically attend in-person classes. Recent literature has focused on providing guidance and discussing the advantages and limitations of online learning, particularly exploring asynchronous and synchronous strategies. These discussions address questions such as when, why, and how to employ these two modes of educational delivery [87]. Specifically, synchronous learning involves a group of participants engaging simultaneously, either in the same physical location (e.g., a classroom) or within the same online environment (e.g., a web conference room). Participants can interact with each other in real-time during these learning experiences. On the other hand, asynchronous learning refers to approaches where instructors and students are not engaged in the learning process at the same time. This can include interactions with pre-recorded

videos or completing on-demand online exams [88].

The increasing focus on synchronous and asynchronous online learning in recent years [89, 90] underscores the importance of analyzing student behavior in these contexts. Such analysis holds significant implications for various aspects, including enhancing teaching strategies, optimizing technological infrastructures, assessing and predicting student engagement, preventing disengagement and potential dropout, and ultimately improving student learning outcomes. Asynchronous learning has been shown to benefit online students by offering flexibility and allowing for more time to reflect on course content. However, asynchronous learning lacks certain aspects that are characteristic of face-to-face lessons, such as the ability to ask questions and the sense of belonging to a class. Therefore, when transitioning from traditional in-person instruction to an online format, synchronous lessons are often preferred in order to replicate some of the features of face-to-face interaction.

The extensive tracking of student behavior in both synchronous and asynchronous learning modes has generated a wealth of data, presenting unprecedented opportunities for the research and educational communities to gain deeper insights into how students learn. While there is a significant body of literature exploring asynchronous online learning [91, 92, 93, 94, 95, 96], particularly focusing on click-stream analysis to predict dropout rates [97, 98] or forecast students' final grades [99, 100], synchronous online learning remains relatively unexplored. Given its increasing adoption, it is crucial to analyze how students learn and model their behavior in this educational modality as well [101]. By conducting in-depth research on synchronous online learning, we can gain valuable insights into student engagement, participation, and learning outcomes in real-time interactive settings.

In this chapter, we investigate how students interact with a synchronous online learning platform in various courses delivered by a university over a semester. Our analysis focuses on collecting and preprocessing students' entry-exit records from lesson rooms, followed by the application of clustering techniques to identify shared behavioral patterns among students at the faculty and course levels. By interpreting and characterizing these clusters, we aim to address the following research questions:

1. **RQ1:** How does the level of student's participation in courses change over a semester?
2. **RQ2:** Which are the principal participation patterns at faculty and course level?
3. **RQ3:** Does the hour of the day a course is delivered influence the level of participation?

4.1.2 Related Work

Our research bridges the gap between two important areas of study: the analysis of synchronous learning from an educational science perspective and the application

of educational data mining in the context of online learning. By combining these two fields, we aim to provide a comprehensive understanding of synchronous learning in online environments.

Synchronous Online Learning

Synchronous online learning is a widely adopted method by teachers for various reasons. It refers to real-time, instructor-led online learning experiences where participants are connected simultaneously and can interact with each other [102]. This mode of learning enables students to ask questions and receive immediate answers, allowing for dynamic discussions and instant feedback. Instructors can assess students' understanding in real-time and adapt their teaching accordingly. The sense of presence and engagement is increased, and interactive activities such as breakout group sessions, live chats, and office hours can be facilitated. Synchronous learning provides a structured schedule that helps students stay on track and promotes task initiation.

The benefits of synchronous online learning have been studied by researchers. For example, Francescucci et al. [103, 104] investigated the effects of a novel synchronous course format, based on virtual interactive rooms, on students' learning outcomes and engagement levels compared to traditional face-to-face instruction. Kohnke and Moorhouse [105] examined the interrelationships between students' perceptions and behaviors in synchronous online learning environments. Recent studies, such as those conducted by Yang et al. [106] and Shoepe et al. [107], have focused on learner behavior during synchronous online lectures, particularly during the COVID-19 pandemic, highlighting the increasing adoption of this mode of learning.

The COVID-19 pandemic has highlighted the challenges faced by educational institutions in managing online learning, including high schools [108]. It has also revealed the vulnerability of the education system to external disruptions [109]. Feldman [110] discusses student assessment during the pandemic and provides recommendations for unbiased and fair grading policies, considering the impact of pandemic-related anxiety, disparities in racial and economic backgrounds, and the need for instructors to adapt to remote instruction effectively.

Overall, synchronous online learning has become a crucial component of educational delivery, particularly in times of crisis. It offers unique opportunities for real-time interaction, engagement, and immediate feedback, while also presenting challenges that require careful consideration and adaptation to ensure equitable and effective learning experiences for all students.

Student Behavior Modeling

Educational data mining focuses on developing methods to explore large-scale data in educational settings [111, 112, 113, 114, 115]. In the context of online learning, several studies have modeled students' behavioral patterns and examined the relationship

between these patterns and learning performance [116]. By analyzing server logs and applying data mining techniques, researchers have identified typical patterns of online learning behaviors and developed predictive models for online learning [117]. For example, Hung et al. [118] applied data mining techniques to analyze student online learning behaviors in a collaborative project-based learning course. Ahuja et al. [119] compared the performance of different clustering and classification algorithms in an educational dataset, while Rodrigues et al. [120] provided a comprehensive review of educational data mining based on clustering in teaching and learning processes.

Clustering approaches have been used to group students based on their learning behavior and personalize the e-learning experience [121, 122]. Other studies have employed clustering algorithms, such as K-Means, to model behavioral patterns of passing and failing students and develop indicators of student performance [123, 124]. These studies have primarily focused on asynchronous online learning.

In our research, we differentiate ourselves by focusing on synchronous online learning, which has received less attention in the field of educational data mining. Although we employ similar techniques, such as K-Means clustering, to model student behavior, our study provides novel observations and contributions specific to the synchronous learning context. We argue that behavioral patterns in synchronous learning are influenced by the unique characteristics of this educational modality. Therefore, our research contributes to a better understanding of student behavior in the synchronous online learning environment.

4.1.3 Methodology

Our research focuses on synchronous learning delivered by the University of Cagliari, an Italian university. The university utilizes Adobe Connect as its e-learning platform, where students can access virtual rooms using computers, tablets, and smartphones. We analyze the data collected from Adobe Connect to understand students' participation levels in synchronous lessons. The dataset includes information on students' participation in synchronous lessons. For each student and each day of a specific university lesson, we analyze their level of participation based on the duration of their connection to the virtual room. By examining this data, we aim to gain insights into students' engagement and behavior in synchronous online learning environments.

University Structure Description. The study encompasses more than 25,000 students across various degree programs. Six faculties are involved in our analysis: *Biology and Pharmacy*, *Engineering and Architecture*, *Medicine and Surgery*, *Science*, *Economic, Law and Political Sciences*, and *Humanities*. Figure 4.1 provides an overview of the number of degree programs offered by each faculty, totaling 89 degree courses. Each faculty designates a coordinator responsible for planning the lessons across programs and booking the virtual rooms in Adobe Connect.

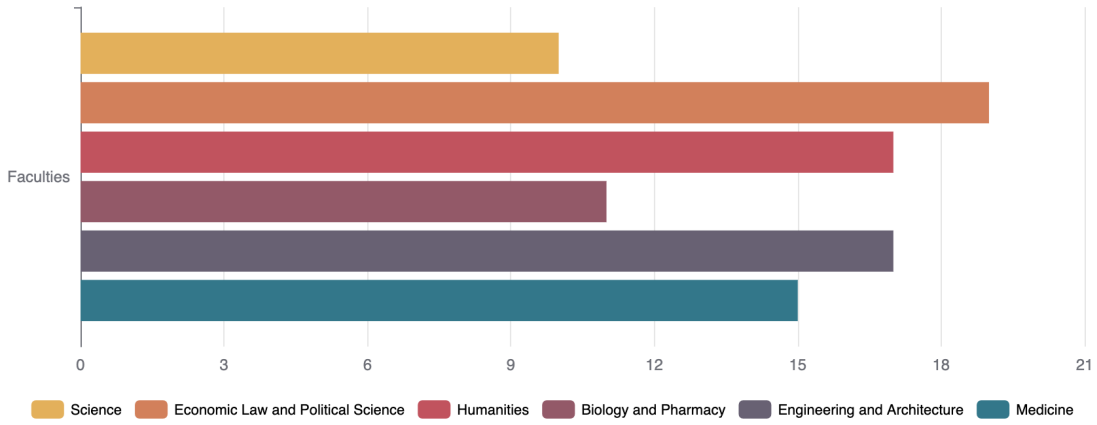


Fig. 4.1: Overview of Bachelor's and Master's Degree Programmes per Faculty.

Online Educational Infrastructure and Delivering. The lessons in our study were conducted using Adobe Connect, a virtual room platform that offers teleconferencing, e-learning sessions, and collaborative content creation and delivery. Each lesson was delivered in a dedicated Adobe Connect virtual room, which served as a permanent online space for instructors to engage with students.

In the context of the University of Cagliari, over 50 virtual rooms were created, with each room assigned to a specific lesson of a course at a designated time slot, similar to classroom assignments in face-to-face settings. Instructors were assigned to virtual rooms during fixed time slots throughout the weeks. Students enrolled in a course could log in using their Adobe Connect accounts and enter the virtual room associated with the desired lesson. Once the instructor initiated the lesson, students could actively participate by raising their virtual hand, engaging in chat discussions, and requesting the activation of video and audio features to ask or answer questions and contribute to the discussion. Figure 4.2 provides a schematic representation of the online synchronous environment used in this study, illustrating the interaction between instructors and students within the Adobe Connect platform.

This online infrastructure provided the necessary tools and features to facilitate real-time interaction and engagement between instructors and students, creating an environment conducive to synchronous learning experiences. In the next sections, we will delve into the analysis of student behavior and participation patterns within this online learning context.

Data Collection Process and Format To collect data on student participation in synchronous lessons, we recorded the entry and exit times of students from the virtual room, as well as the content of the chat, reactions, and raised hands. However, due to privacy considerations, this paper focuses solely on the analysis of entry and exit data. The collected data is organized in a hierarchical folder structure that aligns with the faculty-

degree-course structure of the university. At the top level, there are six sub-folders, one for each faculty. Within each sub-folder, there are Excel files containing multiple sheets, each corresponding to a specific course within the faculty. Each sheet represents the entry and exit records for a particular lesson of that course. Table 4.1 provides an example of the structure of a sheet.

In total, we collected 6,296 sheets of data across the faculties and degree programs, with varying numbers of sheets for each faculty. Specifically, there were 684 sheets for the Faculty of Biology and Pharmacy, 1,320 sheets for the Faculty of Engineering and Architecture, 947 sheets for the Faculty of Medicine and Surgery, 760 sheets for the Faculty of Science, 1,362 sheets for the Faculty of Economic, Law and Political Sciences, and 1,223 sheets for the Faculty of Humanities. This amounted to approximately 464.5 MB of data. The data covers an entire semester, from March to June 2020. The detailed entry and exit data provides valuable insights into student participation patterns and engagement during synchronous online lessons. In the following sections, we will describe our approach to analyzing this data and uncovering meaningful behavioral patterns.

4.1.4 Results and Discussion

In this section, we will analyze student participation patterns throughout the semester, examine the relationship between student groups and the courses they are enrolled in, and explore any correlation between student participation and class schedule. We aim to gain insights into the dynamics of student behavior in the synchronous online learning environment and understand the factors that influence their level of engagement.

Participation Level. We aim to investigate the participation level of students at the faculty level and understand whether the trend of participation changes over the semester. We utilized the entry-exit logs collected in the form of CSV files from Adobe Connect. For each CSV file corresponding to a specific lesson of a course, we counted the number of students present. We excluded cases of university staff logins, empty fields, and

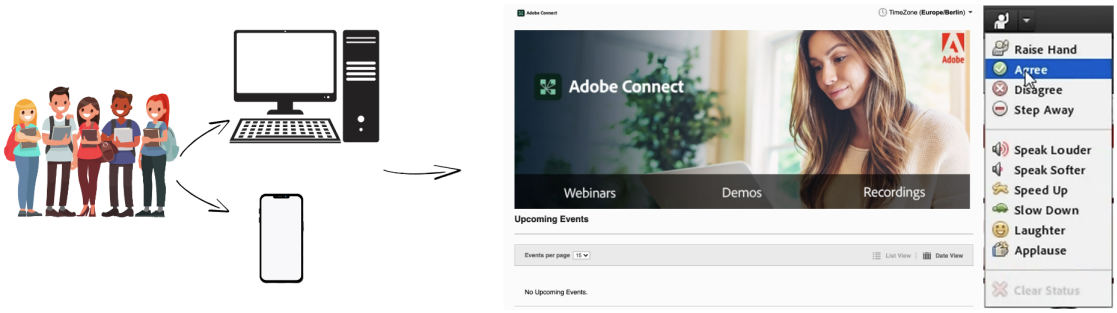


Fig. 4.2: Schematic representation of the online synchronous environment.

Attribute	Description	Example of a registered user (Guest)
transcript-id	The id of the transcription and it is unique for each line	4445626093
asset-id	The id related to that session. When a teacher closes the classroom and ends the meeting, another one with a different id is generated at the next login	4445465743
sco-id	The unique id of the virtual classroom	4211809190
principal-id	The unique id of the registered user. For guest logins this field is empty	3727616656 (-)
login	It is the username and for guests it is empty	student@gmail.com (-)
session-name	It is the name field, which is also present for guests	2019_20/40/12345 John Smith (Guest Name)
sco-name	It is name of the virtual classroom. Now it has changed its name and it is called "training object"	CdL Letters - Room 3
date-created	Is the timestamp of entry into the classroom	2020-11-10 08:37:50
date-end	It is the timestamp of leaving the classroom	2020-11-10 08:38:02 (-)
participant-name	It is the registered username field and it is not present for guests	2019_20/40/12345 John Smith (-)
answered-survey	If a survey is proposed, indicate who answered	0

Table 4.1: Example of data recorded for a lesson of a given course.

duplicates. This procedure was repeated for all lessons within a faculty, and the average number of students present in each lesson was calculated for a given day across all lessons delivered for that faculty on that day.

Figure 4.3 displays the participation level over the semester for each of the six faculties. The x-axis represents the time span from mid-March to early June, and the y-axis represents the average number of students. It is important to note that the participation levels across faculties cannot be directly compared due to the varying number of students in each faculty. Instead, we focus on analyzing the general participation trend within each faculty. The plots indicate a strong initial interest in the lessons, which gradually declines over time. This decline is particularly prominent in the faculties of Medicine and Surgery, Economic, Law and Political Sciences, and Humanities. The Faculty of Science, on the other hand, shows a relatively consistent average number of students throughout the semester, with a slight decline over time. It is worth mentioning that the peaks and dips in the plots are associated with lessons that were planned but not delivered. This highlights the need for a more fine-grained analysis, which we leave for future work.

These participation level patterns provide insights for teachers and support staff. By observing the decreasing number of students over time, teachers can strive to make

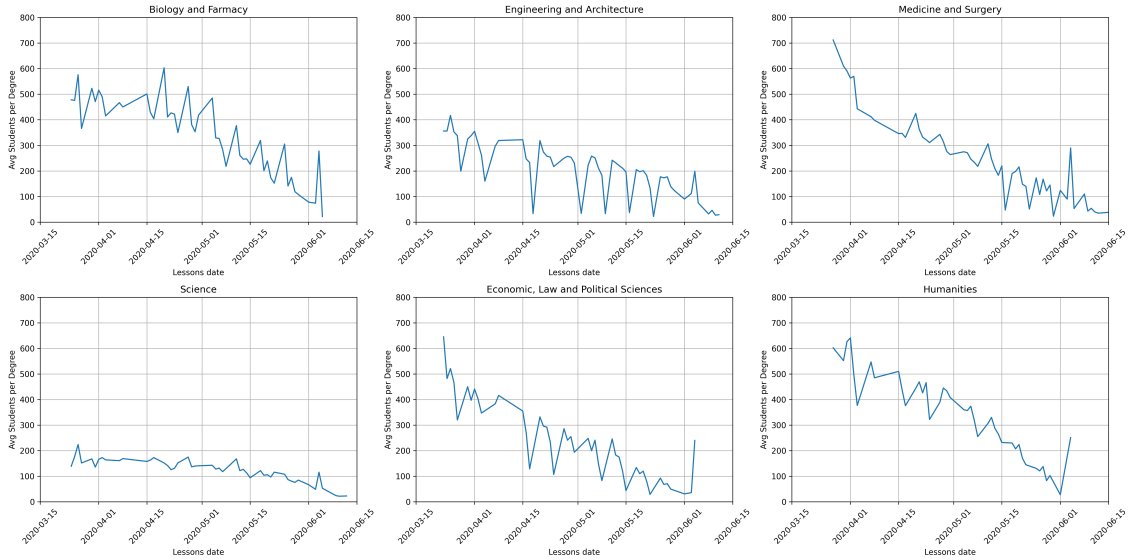


Fig. 4.3: Average number of students per lesson for the six faculties.

lessons more interactive and increase the attention threshold to minimize the drop in student participation. Additionally, these trends suggest the need for adaptive technological infrastructure to accommodate the changing participation levels over time. While these plots provide an overview of the changing participation level over the semester, they do not capture the variation of these patterns within each course of a faculty. Therefore, the next section will further analyze the participation patterns at the course level.

Findings RQ1. *Student participation levels change significantly over a semester, starting high but gradually declining. This trend is consistent across faculties, highlighting the need for ongoing efforts to keep students engaged throughout the term.*

Group Modelling. We investigate the existence of core participation patterns in each course and how these patterns vary across courses. To focus our analysis and better shape our findings, we selected a specific faculty and degree program, namely the *Science* faculty and the *Bachelor's Degree in Computer Science*. We analyzed the second semester, which spanned from March to June 2020, and included 9 courses from the Computer Science program (4 first-year courses, 4 second-year courses, and 1 first-year course). For each course and lesson, we measured the amount of time each student was connected to the platform, representing their participation level. Using this data, we created an N-dimensional vector for each student, where N is the total number of lessons in that course. Each vector contained the minutes the student was connected for each lesson. We then calculated pairwise distances between the vectors and applied the *K-Means* clustering algorithm to the pairwise-distance vectors of all students in each course. To determine the optimal number of clusters, we used the *Elbow* method, using

the Silhouette score as a measure of cluster quality.

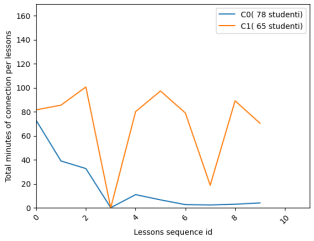
Figure 4.4 displays the centroids of the clusters identified for each course using this methodology. For most courses, two main clusters of students were identified, representing those who consistently followed the course and those whose level of participation decreased over time. This finding is crucial for designing adaptive interventions to motivate students who may need additional support throughout the course. Notably, one course (plot d) exhibited three main clusters. In this case, there is a subset of students who did not engage with the course from the beginning, highlighting the need for multiple levels of adaptive interventions. Additionally, we observed a relationship between the slope of the participation curves for the cluster of students who lost engagement and the time of day for courses delivered during the same year of study. This finding has implications for future lesson scheduling and planning.

It is important to note that some clusters exhibited peaks towards zero in specific lessons. This can be attributed to various factors, such as technical issues or the completion of assigned exercises. These circumstances were observed across multiple courses. Overall, our analysis of group modeling provides insights into the core participation patterns within each course and their variations across courses. Furthermore, considering the time of day in relation to student engagement can inform future lesson scheduling and interventions to support student participation and motivation.

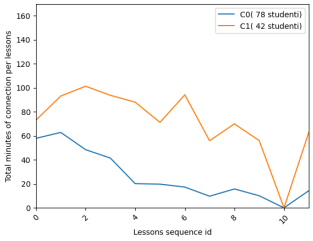
Findings RQ2 and RQ3. *We identified key participation patterns using clustering techniques, with most courses showing two clusters: engaged and disengaging students. Additionally, the time of day affects participation, with earlier courses seeing quicker declines. Educators should use these insights for tailored interventions and optimal scheduling to maintain student engagement.*

Implications and Limitations. In this subsection, we discuss the implications of our findings and acknowledge the limitations of our study. Firstly, it is important to note that our analysis focused solely on the entry and exit data of student connections, without considering intermediate intervals. This means that cases where students were connected for short periods and made multiple intermediate accesses, without actively participating in the lesson for a significant period, were not captured. Incorporating this information could provide a more comprehensive understanding of student engagement during synchronous lessons.

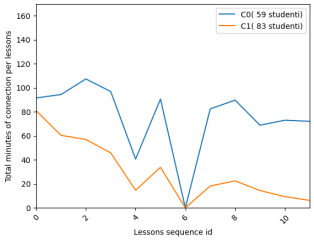
Additionally, due to privacy constraints, we did not consider additional information such as student interventions via chat or system interactions like raised hands. These types of interactions can provide valuable insights into student engagement and participation. Including such data in future studies could enhance the depth of our analysis. Another limitation of our study is the relatively short period of time considered. Our data covers a three-month semester, and extending the analysis to longer periods could reveal different patterns and trends. Examining data from multiple semesters or academic



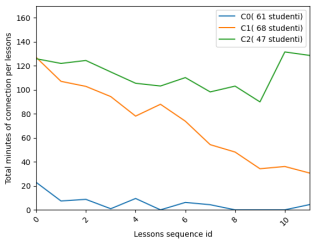
(a) 1st Year
9.00 - 10.40 a.m. (Mon)



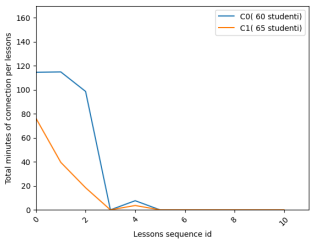
(b) 2nd Year
9.00 - 10.40 a.m. (Wed)



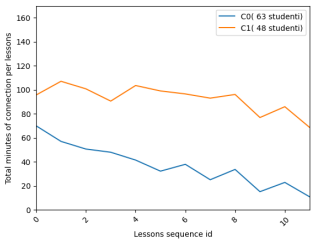
(c) 1st Year
9.00 - 10.40 a.m. (Tue - Thu)



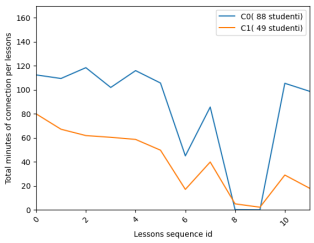
(d) 1st Year
11.00 - 12.40 a.m. (Tue - Thu)



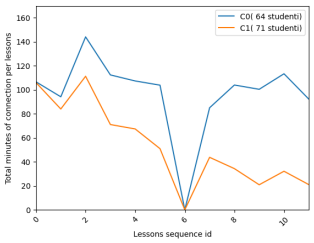
(e) 1st Year
11.00 - 12.40 a.m. (Thu)



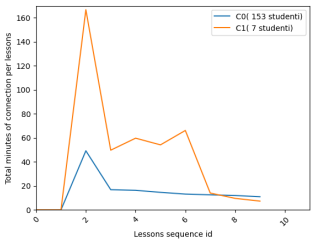
(f) 3rd Year
11.00 - 12.40 a.m. (Mon - Tue)



(g) 2nd Year
3.00 - 4.40 p.m. (Mon - Wed)



(h) 2nd Year
3.00 - 4.40 p.m. (Thu - Fri)



(i) 2nd Year
3.00 - 4.40 p.m. (Tue)

Fig. 4.4: Centroids of the clusters identified for the nine considered courses.

years would provide a more robust understanding of student behavior and participation patterns over time.

Furthermore, our analysis did not account for connection problems that students may have encountered during the lessons. As we have previously observed, these problems can impact the accuracy of the entry-exit records and affect the interpretation of student participation. Exploring and addressing these issues in future research would help to refine our understanding of student behavior in synchronous learning environments. Lastly, we employed the K-Means algorithm for clustering analysis. While this approach provided meaningful insights into student groupings, there are other clustering techniques that could be explored. Comparing different clustering algorithms and their performance on the data would contribute to a more comprehensive analysis.

Despite these limitations, our study has important implications for data-driven support in synchronous teaching. By analyzing student behavior and participation patterns, we can gain valuable insights into student engagement and identify areas where interventions may be necessary. These findings can inform the design of adaptive teaching strategies, personalized interventions, and timely support for students.

4.1.5 Findings and Recommendations

In this work, we conducted an analysis of student behavior in synchronous online learning using data extracted from the Adobe Connect platform. We focused on measuring student participation and modeling their behavior across an entire semester in various courses and faculties of a university. Based on our analysis, we made several key observations and identified potential areas for adaptive interventions and improvements in teaching strategies.

Firstly, we found that the workload on the technological infrastructure varied throughout the semester, with higher workloads observed during the initial period. This highlights the importance of adapting computational resources associated with the platform over time to ensure smooth and efficient delivery of lessons. Furthermore, we observed that a significant portion of students exhibited a decrease in attention and participation as the semester progressed. This trend was particularly evident in certain degree programs. Recognizing this pattern allows for the implementation of adaptive interventions, such as notifying teachers about students who are disengaging or have completely stopped participating. This can help identify the reasons behind the decrease in engagement and enable targeted efforts to keep students engaged throughout the semester. Additionally, we found a correlation between the participation level and the time of day when courses are delivered. Courses held earlier in the day experienced a more rapid decline in student participation. This finding has implications for future lesson scheduling and highlights the need to consider optimal timing for courses to maximize student engagement.

Our work opens up several avenues for future research. Firstly, we plan to expand our analysis by incorporating additional data, including chat interactions, student reactions, and raised hands. This will provide a more comprehensive understanding of student behavior and engagement during synchronous online learning. Furthermore, we aim to compare the participation patterns observed in synchronous online courses with those of face-to-face courses prior to the pandemic. This comparison will help assess the similarities and differences in student behavior across different modes of instruction. Lastly, our findings have practical implications for teachers, as they provide insights into students' participation levels during lessons. Teachers can utilize this information to gauge student attendance and engagement, which can inform their instructional strategies and interventions.

In conclusion, our study contributes to the understanding of student behavior in syn-

chronous online learning and highlights the importance of adaptive interventions and tailored teaching strategies. Through further research and analysis, we can continue to enhance the effectiveness of synchronous online learning and support the success of students in this educational modality.

4.2 Error Analysis on Course Attendance Modelling

In this section, we focus on predicting the attendance of students in synchronous online courses using machine learning techniques. By analyzing student participation logs and extracting relevant features, we aim to develop models that can predict the level of student attendance throughout the course. The prediction of course attendance can provide valuable insights for instructors and course managers, allowing them to identify students at risk of disengagement and implement timely interventions to improve attendance and overall student success.

4.2.1 Introduction

Context. With the increasing adoption of digital technologies, many universities have expanded their offerings to include synchronous online courses, where participants engage in real-time, instructor-led learning experiences. Synchronous learning provides a dynamic environment where students can actively participate, ask questions, and receive immediate feedback [102]. As such, it has become a widely adopted teaching modality in higher education institutions. Numerous studies have explored the effectiveness and engagement levels of students in synchronous online lectures, comparing them to traditional face-to-face formats [103, 104, 106, 107].

Ensuring the attendance and quality of instruction in synchronous courses is crucial for supporting higher education systems in meeting emerging needs and challenges while prioritizing student qualifications and experience [125]. Universities employ periodic evaluation processes to improve educational quality and increase attendance. These processes often form part of a formal quality assurance model, incorporating internal procedures and external checks from third-party agencies. As higher education continues to evolve, it requires a student-centered approach, flexible learning paths, and continuous refinement of courses to meet diverse expectations [126].

Open Issues. Monitoring course attendance and quality is an important yet challenging task. Traditional processes rely on end-of-semester questionnaires to gather students' opinions. The collected answers are then analyzed by didactic managers and discussed with individual instructors and the board of instructors of the respective degree program. However, this evaluation process occurs after the semester has ended, resulting in a time gap between data collection and feedback provision. Consequently, instructors have limited opportunities to make interventions during the current semester and can only

apply improvements for subsequent iterations. Ideally, these procedures should provide timely insights to support instructors in refining their teaching methods and enhancing student learning outcomes.

To ensure that students in the current iteration receive the necessary support, monitoring course quality and attendance throughout the semester is essential. However, frequent collection of attendance forms and quality questionnaires from students, as well as the subsequent analysis by didactic managers and instructors, is time-consuming and not scalable. To address this limitation, recent research has explored the use of machine learning techniques for predicting course quality and attendance. Initial steps have been taken in this direction [127, 128, 129]. For example, one study modeled student behavior through attendance records from a large university [130]. However, these approaches have focused on a narrow set of courses and relied on video interaction logs that are not available for synchronous courses. Therefore, there is still ample room for research on synchronous course attendance and quality prediction.

Our Contribution. In this paper, we investigate whether student attendance patterns up to a certain lecture can predict the quality and future attendance of synchronous courses within an online real-time classroom scenario. To accomplish this, we preprocess both the student participation logs for all the courses provided by a public university and the quality indicators provided by students in the final questionnaires for those courses. We then extract predictive features from the behavioral patterns of student participation and employ a machine learning approach to predict future attendance and quality indicators based on these features. Finally, we discuss our results and the main implications for course attendance and quality prediction.

4.2.2 Methodology

Our methodology aims to assess the predictive power of machine learning models for early course attendance and quality prediction, utilizing patterns extracted from past attendance logs. We implemented a supervised classification pipeline to accomplish this task.

Firstly, we collected student participation logs from the online learning platform Adobe Connect, which was used for delivering synchronous courses in our context. These logs included information such as the course ID, lesson ID, student ID, and the timestamps indicating the entry and exit times of students from each lecture. Next, we extracted a range of relevant features from these attendance records to capture various aspects of student participation behavior. These features encompassed course-level properties, such as the time of day the lectures were delivered, as well as lecture-level characteristics, including the average time spent by students in a particular lecture and the proportion of the lecture attended by each student. Additionally, student-level features were considered, such as the number of courses a student was attending and their tendency to join lectures late or leave early. Subsequently, we trained and evaluated

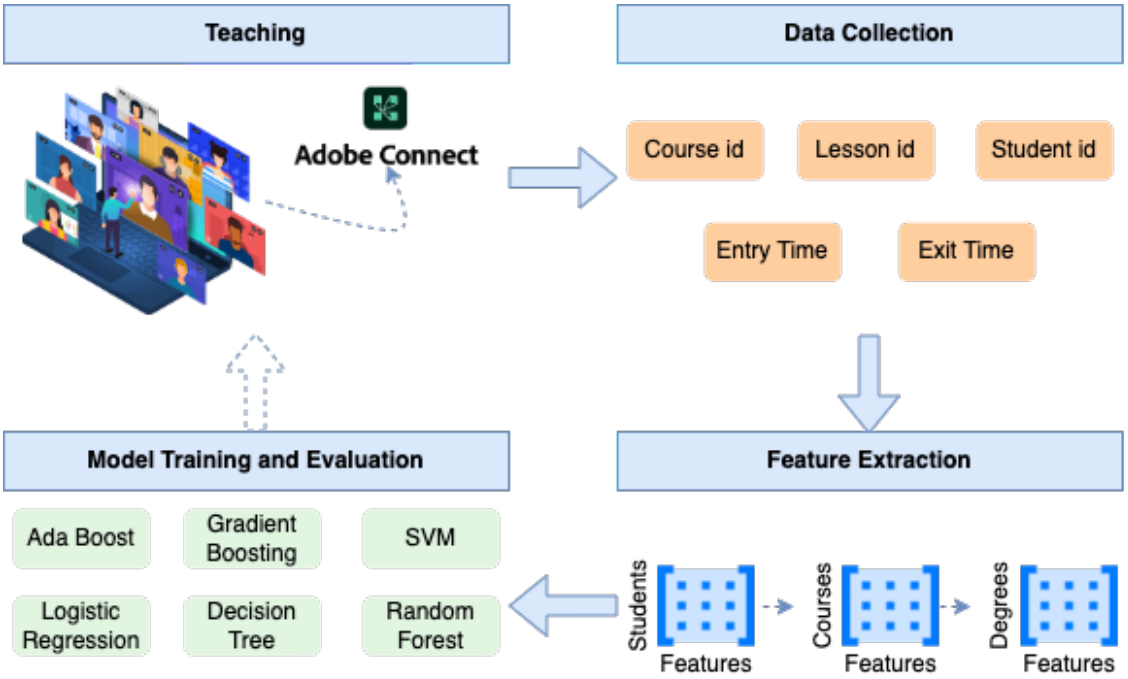


Fig. 4.5: Methodology overview: Collecting student participation logs, extracting relevant features, and training classifiers for predicting course attendance and quality.

classifiers separately for predicting future course attendance and quality indicators. Our classification models were trained using the extracted features as input and the corresponding attendance or quality indicator as the target variable. We employed a variety of machine learning algorithms, including Random Forest, to train the classifiers. The evaluation process involved a nested stratified 10-fold cross-validation, and we utilized the balanced accuracy metric to assess the performance of the models.

The overall framework of our methodology is illustrated in Figure 4.5, which provides a visual representation of the various stages involved in the process, from data collection to model training and evaluation. By leveraging this approach, we aim to gain insights into the early prediction of course attendance and quality, enabling instructors and course managers to make informed decisions and interventions to enhance the overall educational experience for students.

Data Collection

For our study, we collected student attendance logs from a large public university encompassing over 25,000 students, 6 faculties, 89 degree programs, and 1,230 courses. The data collection process adhered to strict privacy and ethical considerations, ensuring the anonymity and confidentiality of the participants.

Table 4.2: Schema of the data structure and fields leveraged in this study.

Entity	List of Attributes	#Records
Degree	Degree id, level (BSc, MSc).	89
Course	Course id, year, degree id.	1,230
Lessons	Lesson id, teaching id, date, start time, end time.	13,000
Accesses	Lesson id, student id, access time, exit time.	525,000
Students	Student id, year of attendance, course id.	25,000
Indicators	Teaching id, question category, question criterion, overall grade.	3,500

In this university, each lecture of a course was delivered synchronously in a virtual room, facilitated by an online learning platform. Students enrolled in a course could log in to the virtual room and access the specific virtual room associated with the lecture they were required to attend. Throughout an entire semester, we tracked the entry and exit times of students from the virtual room for each lecture. This information was recorded in the attendance logs, which included the course ID, lecture ID, student ID, entry timestamp, and exit timestamp. It is important to note that similar attendance logs can also be collected for in-person face-to-face lectures, albeit with different mechanisms.

Table 4.2 provides a comprehensive overview of the collected data, including the number of students, faculties, degree programs, and courses involved in our study.

The availability of this comprehensive dataset allows us to analyze and evaluate student attendance patterns and their predictive power for course quality indicators, providing valuable insights for educational stakeholders to improve the teaching and learning experience.

We also collected the quality indicators for all degree programs offered during the semester. These indicators were computed based on the students' responses to the university questionnaire, which evaluated various aspects of course quality. However, for privacy reasons, we were unable to access the specific quality indicators computed for individual courses within each degree program.

As a result, our study focused on predicting the overall quality of degree programs rather than individual courses. For each degree program, we computed 14 course quality indicators that covered various stages of the course, including pre-course, in-course, and post-course indicators. These indicators encompassed aspects such as preliminary knowledge, study workload, course material, examination method, content novelty, punctuality, motivation, clarity, tutoring activities, syllabus coherence, availability, lecture interest, overall satisfaction, and online satisfaction. Each quality indicator was measured on a scale ranging from AA to F, as shown in Table 4.3. To facilitate our preliminary experiments, we binarized the indicator values by assigning a label of 0 to a degree program if the indicator value fell below the average value across all degree pro-

Table 4.3: Levels scale for each considered quality indicator.

Level	Description
AA	Very positive
A	Overall positive, situation to be consolidated
B	Sufficiently positive, situation with room for improvement
C	Slightly positive, situation with considerable room for improvement
D	Slightly critical, attention required
DD	Critical, intervention is required
E	Very critical, intervention is particularly required
F	Extremely critical, structural intervention required

grams. Conversely, a label of 1 was assigned if the indicator value exceeded the average. By categorizing the quality indicators in this manner, we aimed to explore the predictive power of student attendance patterns in relation to the overall quality of degree programs.

Feature Extraction

The relationship between students' behavioral aspects and academic achievement has been widely studied [131, 62, 63]. Important dimensions of learning behavior include persistence, effective time management, self-awareness, and careful examination of course materials. Previous studies, such as [132], have proposed various sets of features based on these dimensions. However, these studies focused on asynchronous courses that utilized pre-recorded microlearning videos. In our study, we aimed to derive features with a similar rationale but adapted to the available logs and synchronous learning scenarios.

We extracted a range of features for each course, as shown in Table 4.4. These features capture different levels of the learning environment. Course-level features include properties that may influence students' perceived quality of the course, such as the time of day the course lectures were delivered and the distribution of the number of students attending the lectures. Lecture-level features capture students' participation behavior within each lecture, including the average time spent by students in the lecture, the average proportion of the lecture followed by students, and the number of students who attended the lecture. Student-level features are related to individual student characteristics that can influence their attendance, such as the number of courses the student is attending and their tendency to join late or leave early during a lecture.

These features were extracted from the raw participation logs. Since we only had access to quality indicators at the degree program level, when predicting course quality indicators, we averaged the features of all students and lectures within a course. Similarly, we averaged the features of all courses within a degree program to obtain a single feature vector per degree program.

Table 4.4: Features extracted from attendance logs in our study.

Dimension	Feature Name	Description
Student level	Late	The student logged in after class time
	Hasty	The student logged out earlier than the average time of the exits
Lecture level	Attendance rate	Percentage of the lesson attended
	Avg jump	Average of "jumps" related to the access time between one lesson and the previous one
	Average time access	The average time spent per lesson
	Standard deviation	The standard deviation of the time spent in each lesson
	Max and min time	The highest and lowest access time of a student
	Daily lessons	Number of lessons attended on the same day
	Lessons per day	Number of daily lessons taken by a student over time using Kurtosis and Skewness
	Lesson time	Average time of attended lessons
Course level	Teaching participation	Measurement of the level of participation through Kurtosis and Skewness
	Attended courses	The number of courses attended by the student in the same period

Model Creation

In our study, we considered a variety of machine learning models that have shown high accuracy in education-related scenarios, even though our focus was on a different prediction task. The following models were included in our analysis:

- **AdaBoost Classifier:** An adaptive classifier that can improve the performance of other learning algorithms. It is commonly used for binary classification but can be extended to multiple classes or limited ranges on the real number line.
- **Gradient Boosting Classifier:** This classifier combines the AdaBoost method with minimization techniques to minimize the difference between the actual and expected class values in the training examples.
- **Support Vector Machines (SVM):** A classification technique originally designed for binary classification but can be extended to handle multi-class problems by decomposing them into a series of binary sub-problems.
- **Logistic Regression Classifier:** It uses a logistic function to model the dependent variable, making it suitable for binary data classification tasks.
- **Decision Tree Classifier:** This technique represents a set of classification rules in a tree structure using the "if-then" format.
- **Random Forest Classifier:** An ensemble technique that combines multiple decision trees to create a forest, reducing overfitting and improving the predictive accuracy.

While we explored various models, for the sake of conciseness, we will primarily report the performance of the Random Forest Classifier. This model has often demonstrated a good balance between prediction accuracy and interpretability of the results.

Model Evaluation

To evaluate the performance of our models, we employed a nested stratified 10-fold cross-validation approach. The purpose of this approach was to ensure that the folds were divided by course/degree, maintaining the integrity of the data structure. We applied the same folds for all experiments across the models. During the evaluation process, we optimized the hyperparameters using grid search. In each iteration of the outer cross-validation loop, we performed an inner 10-fold cross-validation on the training set. This allowed us to select the combination of hyperparameter values that yielded the highest accuracy on the inner cross-validation.

To assess the validity of the models, we computed the Area under the ROC Curve (AUC) on the training and validation sets. This measure provided insights into the models' performance on both the training and unseen validation data. Additionally, we evaluated the generalizability of the models by computing the AUC on the test set. It's important to note that our models were trained on a per-lecture basis. This means that the model for a specific lecture l of a given course was trained using features extracted from data collected up to and including lecture l. This approach allowed us to capture the temporal dynamics of student behavior and make predictions based on the available information up to a given point in time.

4.2.3 Experimental Results

In this subsection, we present the experimental results of our study on course attendance and quality prediction using attendance records. The models, evaluation results, and source code are available for replication on GitHub. Our methodology is flexible and can be adapted to analyze various perspectives, but we primarily focused on addressing the following research questions:

- **RQ1:** Can we accurately and early predict course quality indicators given by students to a degree program?
- **RQ2:** To what extent are our extracted features and models predictive of whether a student will attend the next lecture of a course?
- **RQ3:** Can we early predict whether a student will attend a certain number of the subsequent lectures?

By exploring these research questions, we gained insights into the predictive power of attendance records and extracted features in relation to course quality indicators and student attendance behavior.

Course Quality Prediction (RQ1)

In our initial analysis, we examined whether student participation features could predict the aggregated quality indicators of a degree program early in the course. Figure

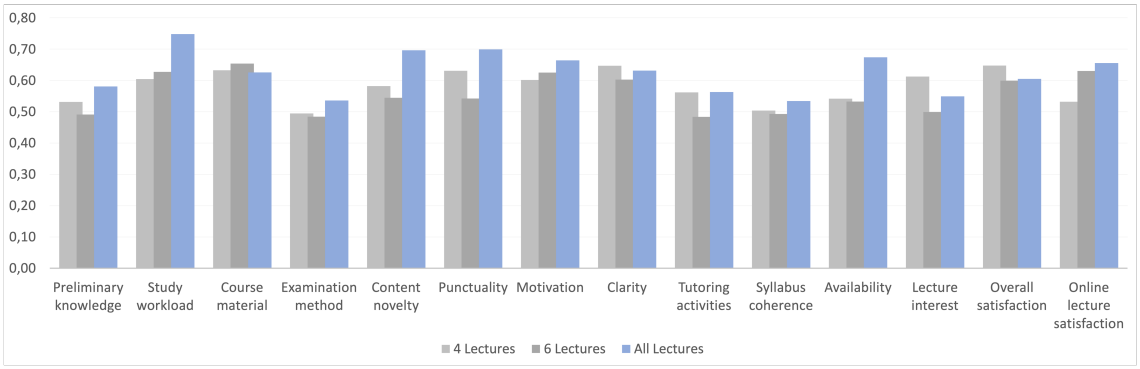


Fig. 4.6: Models performance (AUC) on the course quality prediction task (RQ1).

4.6 illustrates the performance of the Random Forest Classifier trained on attendance records up to the fourth and sixth lecture (gray bars), as well as the model trained on all lectures (blue bars). The results showed that when using the full course data, the prediction performance varied across different quality indicators. The Area under the ROC Curve (AUC) ranged from 52% for the examination method indicator to 76% for the study workload indicator. Notably, overall satisfaction, including online activities, exhibited reasonable predictability.

Indicators related to the course content, such as study workload, course material, and clarity, along with instructor-related indicators like punctuality, motivation, and availability, demonstrated relatively good predictability. However, content novelty, punctuality, and availability could only be predicted with reasonably good accuracy when using the full course data. Interestingly, the prediction of lecture interest performed better after four lectures compared to the full course. These findings suggest that not all quality indicators can be accurately predicted with a reasonably good accuracy, even when using the full course data. However, for six out of the nine indicators that were predictably accurate with the full course data, reasonably good predictions could be made as early as after six lectures, which corresponds to a few weeks of the course. It is worth noting that the attendance records may not provide sufficient information to predict indicators that involve more complex aspects. Future work should consider incorporating fine-grained logs of activities and interactions during lectures to enhance the monitoring and tracking of learning in synchronous courses.

Findings RQ1. *Not all indicators could be predicted with a reasonably good accuracy of at least 63-65%, even when the full data was available. However, indicators that exhibited predictability, even after six lectures, showed performance close to that of models trained on all lectures.*

Fig. 4.7: Models performance (AUC) on the next lesson participation task (RQ2).

Predicting Attendance to the Next Lecture (RQ2)

In our second analysis, we investigated the predictive capability of different models to determine whether a student would attend the next lecture based on their behavior up to a certain lecture. The AUC scores for all models, using features extracted until the fourth, sixth, and eighth lecture, are presented in Figure 4.7.

The Random Forest Classifier demonstrated the highest performance among the models. After only four lectures, this classifier achieved an AUC of 84%. The AUC score further increased to 89% after six and eight lectures. The Gradient Boosting Classifier was the second most accurate model, with AUC scores of 85% after four and six lectures, and 87% after eight lectures. Notably, the performance gap between the Random Forest Classifier and the Gradient Boosting Classifier was relatively small. However, the Support Vector Machines model performed the worst, with AUC scores ranging between 56% and 58% after four, six, and eight lectures, respectively.

Overall, with the exception of the Support Vector Machines model, all other models demonstrated reasonably accurate predictions regarding student attendance to the next lecture. The Random Forest Classifier exhibited the highest accuracy in this study. Although performance was already high after only four lectures, models trained after six and eight lectures showed more stable behavior.

Findings RQ2. *With the exception of the Support Vector Machines model, all models provided reasonably accurate predictions regarding student attendance to the next course lecture. The Random Forest Classifier demonstrated the highest accuracy in this study.*

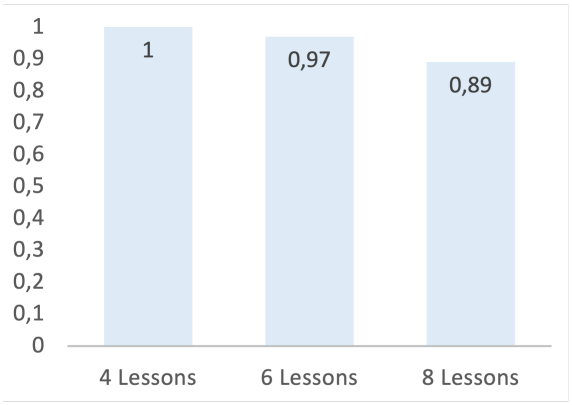


Fig. 4.8: Percentage of students for whom the course attendance requirement prediction was not trivial (RQ3).

Predicting Attendance to a Certain Percentage of Course Lectures (RQ3)

In our third analysis, we aimed to investigate whether it is possible to predict, using machine learning models, whether a student will meet the course attendance requirement typically imposed by universities. Many universities require students to attend at least a certain percentage of the course in order to be eligible to take the final exam. In our experiments, we set the threshold at 70% of course lectures.

Before training models for this task, we performed an exploratory analysis to determine the number of students for whom the prediction task would be non-trivial. For instance, if a student missed the first four lectures of a 10-lecture course, it would be obvious that they would not meet the attendance requirement. Therefore, we excluded those students from our analysis. Figure 4.8 illustrates the percentage of students for

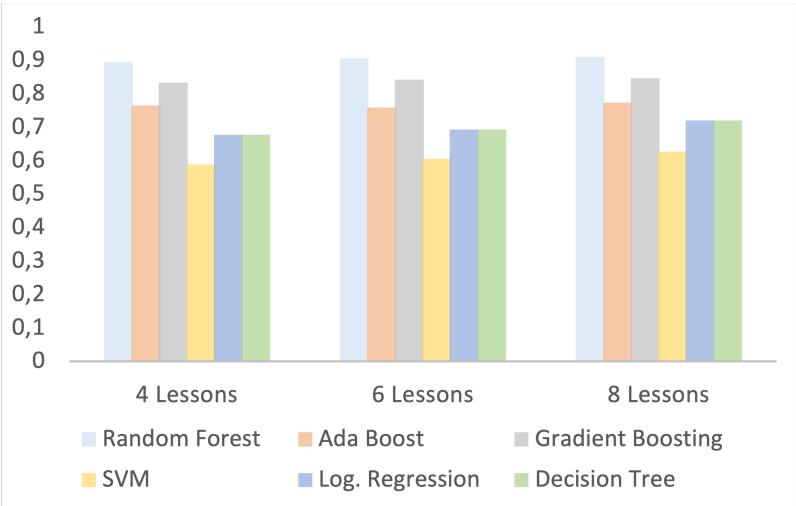


Fig. 4.9: Performance in terms of AUC for the course requirement task (RQ3).

whom predicting their participation in at least 70% of the lectures was not trivial. After four lectures, all students were included in the analysis. However, there was a 3% decrease after six lectures, and by the eighth lecture, the percentage of students considered dropped to 89%. For the remaining students, we trained the models and evaluated their performance in terms of AUC, as shown in Figure 4.9. Each model exhibited distinct performance patterns. The Random Forest Classifier demonstrated stable performance across lectures, with only a marginal increase in AUC. After four lectures, the AUC reached 89%, and it increased to 90% and 91% after six and eight lectures, respectively. This model proved to be effective for predicting whether a student would meet the attendance requirement. The Ada Boost Classifier exhibited consistent performance, with AUC scores ranging from 75% to 77% across all settings. Although reasonably high, the AUC scores for this model were lower than those of the Random Forest Classifier. The Gradient Boosting variant performed better, with AUC scores ranging from 83% to 84%. This model outperformed the others after four, six, and eight lectures. It is important to note that Gradient Boosting is more efficient than Random Forest and should be preferred for large datasets or when computational resources are limited. On the other hand, Support Vector Machines reported the lowest AUC performance, ranging from 58% to 62%. These scores were close to the AUC value of 50%, which indicates random prediction. Therefore, Support Vector Machines are not suitable for this prediction task. Logistic Regression and Decision Trees exhibited similar and relatively low performance, with AUC values ranging from 67% to 72%.

Findings RQ3. *Not all models were able to accurately predict whether a student would meet the course attendance requirement of participating in at least 70% of the course lectures. The Random Forest Classifier demonstrated the best performance in terms of AUC, followed by Gradient Boosting. The other classifiers did not yield reasonably high performance estimates for this task.*

4.2.4 Findings and Recommendations

In this study, we utilized a machine-learning pipeline to predict course quality and student attendance based on student participation data. Our approach offered a scalable and transparent alternative to manual practices that rely on questionnaires. We examined various dimensions of student participation across a diverse range of courses in a public university.

Our findings demonstrate the predictive power of student attendance behavior and patterns. By analyzing the data collected from the first four lectures, we were able to accurately predict several quality indicators, the likelihood of attending the next lecture, and the probability of attending at least 70% of the lectures. This information can be valuable for instructors to improve their course delivery and take proactive measures to enhance students' academic experience from the early stages of the course. For example, instructors can plan activities based on the expected number of attendees for the next lec-

ture. Our models empower instructors to identify and address low class participation and course quality issues. Furthermore, our models can be applied to both online distance education and face-to-face settings where digital entry and exit systems are in place. However, it is essential to conduct user studies involving students and instructors to assess the effectiveness of the system in real-world scenarios. This feedback will inform future improvements and enhancements to the implementation of the system, ensuring its accuracy and reliability. Moving forward, we plan to extend our analysis to predict course quality at the individual course level, as our current predictions are limited to degree program level due to privacy constraints. We also aim to explore different contexts, such as other universities, and investigate alternative features and predictive models. Additionally, if the system predicts low class participation for a significant number of students in a particular course or degree program, it could inform the implementation of specific support services, such as supplementary courses or personalized guidance.

By continuously collecting and analyzing data in a systematic and consistent manner, we envision the development of a comprehensive tool that integrates predictive models and provides valuable insights to stakeholders. We intend to further examine the predictiveness across different faculties, study levels, and teaching modalities and incorporate model predictions into interactive dashboards. This holistic approach will contribute to the enhancement of course quality and student engagement in higher education settings.

Chapter 5

Analyses of Algorithmic Performance in Asynchronous Learning

In Chapter 4, we ventured into synchronous learning environments, examining algorithmic performance and predictions. Now, in this Chapter, our focus shifts to asynchronous learning environments. This chapter delves deep into the analysis of student behavior, exploring participation patterns and their implications. Our primary aim is to identify trends and insights that can inform and improve teaching strategies, optimize technological infrastructures, and enhance overall student learning outcomes.

5.1 Error Analysis on Student Success Prediction

5.1.1 Introduction

Context and Objective. Learning analytics has emerged as a promising field that leverages data and analytical studies to understand factors influencing student success and provide recommendations for improving the learning process [133]. Predictive models have been developed to forecast student outcomes based on various indicators such as engagement, regularity, critical thinking, metacognition, and socio-emotional well-being [131, 134, 135, 136, 137, 138, 139, 140, 141]. These models have proven useful for personalized interventions, adaptive content delivery, and understanding the learning process in instructional strategies like blended learning and online learning [142, 143]. However, the adoption of these predictive models also introduces risks for both students and instructors. Users of these models may not be accustomed to reasoning about model uncertainty, which can lead to misunderstandings or mistrust in the predictions [144, 145, 146]. Trust is crucial for the acceptance of these models in educational settings. Therefore, it is important to investigate the weak spots of student success models and understand the circumstances under which these models should (not) be trusted.

Open Problem. One particular weak spot that has not been extensively explored in

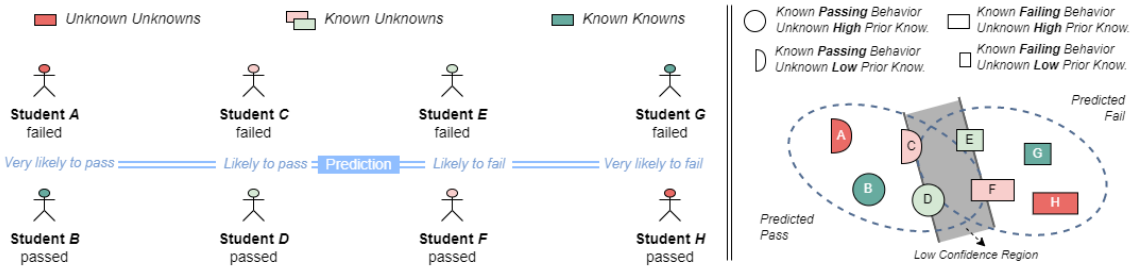


Fig. 5.1: Motivating Example for the Unknown Unknowns Problem.

student success models is the presence of "unknown unknowns" [147]. Unknown unknowns refer to instances where the model is highly confident in its prediction despite being incorrect. These instances arise when the model lacks knowledge of the correct labels (model knowledge) and is not aware of its own incorrect predictions (model awareness). Unknown unknowns can be attributed to various factors such as unmodeled data biases, data distribution shifts, or hidden factors not captured by the model, such as students' prior knowledge. While efforts have been made to analyze unknown unknowns in fields outside of education, their impact on student success prediction models remains unexplored [148, 149, 150].

Motivating Example. To illustrate the unknown unknowns problem, consider the following example depicted in Figure 5.1. On the left, we see a student success model highly confident, but actually wrong, on predictions for students A and H (unknown unknowns), both highly confident and correct on students B and G (known knowns), and only slightly confident of its predictions on students C, D, E, and F (known unknowns). On the right, we assume that students' prior knowledge was a variable the model was not aware of. In particular, a first-year university student, referred to as Student A, failed a Math course taught in a flipped format. During the course, students had access to a platform for pre-class activities, including watching videos and doing exercises. A machine-learning model was employed to identify students in need of intervention based on their pre-class behavior. The model, trained on historical data, demonstrated high accuracy but was unaware of the majority of students' high prior knowledge in Math. Consequently, the model erroneously associated a low time spent on videos with success, as the content was often confirmatory knowledge. Student A struggled with the course and did not spend much time on videos. Despite being at risk, the model predicted that Student A would likely pass the course. Due to limited teaching resources, no intervention was provided to Student A, who eventually failed the course. In this case, unknown unknowns resulted from students' prior knowledge that the model was unaware of. We illustrate that not only students' behavior (visible to the model), but also students' prior knowledge (outside of the scope of the model) was relevant for success. Student A exhibited a typical passing behavior, but had a low prior knowledge. On the contrary, Student H exhibited a typical failing behavior, but had a high prior knowledge.

Our Contribution. In this work, we aim to uncover and characterize unknown unknowns in student success predictions. We investigate the existence of unknown unknowns in student success prediction models and explore their variations across different instructional settings, such as flipped courses and Massive Open Online Courses (MOOCs). Furthermore, we propose a computational framework to identify and characterize unknown unknowns, and we assess its effectiveness, informativeness, and cost-efficiency using state-of-the-art student success prediction models. Our analysis is based on data collected from six courses, including three flipped classroom courses and three MOOCs. Through our experiment, we investigate three key aspects:

1. We examine the existence of unknown unknowns in student success prediction and explore how their prevalence and types vary across different learning environments, such as flipped courses and MOOCs.
2. We explore the feasibility of characterizing unknown unknowns in student success prediction models, shedding light on the factors contributing to their emergence.
3. We assess the impact of providing instructors with information about unknown unknowns on their perception of the student success model, aiming to understand how this knowledge influences instructional practices.

Our results reveal that unknown unknowns pose a significant challenge for student success models, underscoring the importance of our framework in identifying and understanding these elusive cases across various experimental conditions. This work contributes to the field by shedding light on the complexities of student success prediction and highlights the need for more nuanced approaches that consider unknown unknowns.

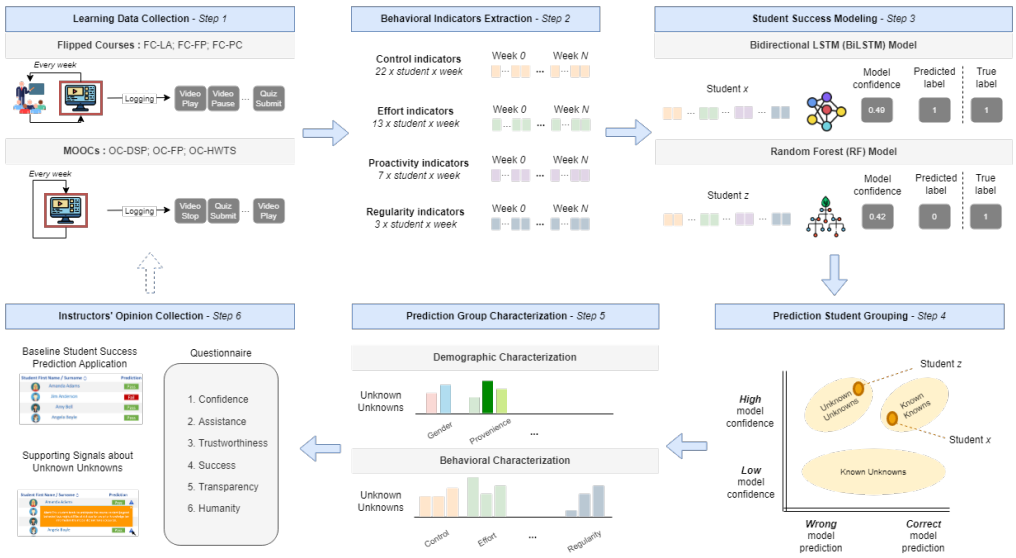


Fig. 5.2: Methodology steps for detecting and characterizing students performance

5.1.2 Methodology

In this subsection, we outline the methodology employed in our study, including the educational scenario, data collection process, behavioral indicators, student success models, and the overall analysis framework. Figure 5.2 provides an overview of the different steps involved in our framework. We collected log data from flipped courses and MOOCs (Step 1) and extracted behavioral indicators of control, effort, proactivity, and regularity (Step 2). We built a random forest classifier and a bidirectional LSTM classifier, both returning a confidence score and a predicted label (Step 3). We then grouped students into known knowns, known unknowns, and unknown unknowns based on the confidence score, predicted label, and original true label (Step 4). The three groups were characterized to identify (dis)similarities (Step 5). Finally, we asked instructors for their opinion and support on unknown unknowns detection (Step 6).

Learning Data Collection - Step 1/6

Our study specifically focused on two teaching strategies implemented in courses at a European university: flipped classroom courses and Massive Open Online Courses (MOOCs). These teaching scenarios were chosen as they represent distinct approaches to instruction within an online-based learning environment. In flipped classroom courses, students engage in pre-class activities, such as watching instructional videos, reading materials, or completing assignments, before attending face-to-face sessions. The pre-class activities aim to provide students with foundational knowledge and prepare them for in-person discussions and collaborative activities. MOOCs, on the other hand, are fully online courses that are accessible to a large number of participants. They typically consist of video lectures, interactive exercises, quizzes, and discussion forums. MOOCs often attract a diverse range of learners from around the world and offer flexible learning opportunities. For both flipped classroom courses and MOOCs, we collected data on various aspects of student engagement and behavior, including the duration of video views, completion rates of pre-class activities or course modules, participation in discussion forums, and performance on quizzes or assignments. These data points were instrumental in understanding the learning patterns and behaviors of students within these different instructional contexts. By focusing on these two teaching strategies, we aimed to capture a wide range of learning activities and behaviors that could potentially influence student success and contribute to the identification of unknown unknowns in student success prediction.

Learning through Flipped Classroom Courses. Our analysis focused on three semester-long university courses that followed the principles of learning science outlined in [152]. These courses were mandatory for students pursuing Bachelor's degrees in Computer Science and Communication Systems at the European university under study. The details of these courses are provided in Table 5.1. The instructional format of these flipped classroom courses included a combination of lectures and recita-

Table 5.1: Detailed information about the courses.

Course Title	ID	Field[1]	Setting	Students[2]	Level	Language	Weeks	Failing Rate	Quizzes
Linear Algebra	FC-LA	Math	Flipped	292	BSc	English	14	40.00%	179
Functional Programming	FC-FP	CS	Flipped	216	BSc	French	18	38.53%	0
Parallelism and Concurrency	FP-PC	CS	Flipped	147	MSc	French	16	37.16%	0
Digital Signal Processing	MOOC-DSP	CS	MOOC	15,394	MSc	English	10	75.71%	38
Household Water Treatment and Storage	MOOC-HWTS	NS	MOOC	2,423	BSc	French	6	47.36%	10
Functional Programming Principles in Scala	MOOC-FP	CS	MOOC	18,702	BSc	French	8	42.15%	3

[1] **Field**: *CS*: Computer Science; *Math*: Mathematics; *NS*: Natural Science. [2] **Students**: for MOOCs, number of students obtained after removing early-dropout students [151].

tion or exercise sessions on a weekly basis. In-class activities involved quizzes, short problem-solving exercises, and structured proof-type problems. Additionally, students were expected to spend a few hours each week on individual study as part of their pre-class activities. One week prior to each class, students received instructions regarding the preparatory work, which consisted of a list of sections from a Massive Open Online Course (MOOC) containing video lectures and online quizzes. The quizzes in the MOOC allowed students to self-assess their learning progress. Data pertaining to student pre-class activities were collected by the MOOC platform for all three flipped courses. The logged entries included information such as the user ID, specific activity (e.g., playing a video), and timestamp (e.g., date and time of the activity). In addition to activity data, demographic attributes of the students, including gender, geographic origin, and high school diploma, were also recorded. However, no data was collected on in-class activities. To assess student achievement, the final exam grades were used, with a passing grade defined as 4 or higher on a scale of 1 to 6. The study received ethical approval from the university's ethics committee (HREC 058-2020/10.09.2020, 096-2020/09.04.2020).

Learning through Massive Open Online Courses. To complement our analysis, we included three Massive Open Online Courses (MOOCs) taught by three different instructors from the same European university on the Coursera platform (as shown in Table 5.1, last three rows). These courses were accessible to learners worldwide. The MOOCs followed a weekly release format, where new lecture content was made available each week. Students were expected to dedicate several hours per week to complete the course materials. Each week, the courses consisted of short video lectures, typically ranging from 10 to 15 minutes, introducing key concepts, followed by quizzes for self-assessment. Additionally, students were required to complete graded assignments on a weekly basis. The instructor used these assignment scores to evaluate student achievement. The final course grade was calculated by weighting the scores from the weekly assignments and the final exam, with a passing grade requiring a minimum of 60 points on a scale of 0 to 100. For the MOOCs, we collected a total of over 145,640 log entries, which included information such as the user ID, specific activity, and timestamp. The format of the log entries was consistent with that of the flipped courses. Students' gender and geographic origin were also attached to the log data when voluntarily provided by the students. The collected data encompassed all the activities performed by

the students on the MOOC platform, representing a substantial portion of their learning experience (excluding offline activities such as video watching). In the MOOC setting, our study focused on scenarios where student success models were built using the entire activity data available on the platform, which was not visible to the instructors. Student achievement in the MOOCs was measured based on the final course grade.

Behavioral Indicators Extraction - Step 2 / 6

Previous studies have highlighted the significant association between academic achievement and various aspects of self-regulated learning (SRL), such as effort regulation, time management, metacognition, critical thinking, and help-seeking [131, 62, 63]. These aspects have been the focus of different learning indicator sets proposed in the literature. The effectiveness of these indicators in modeling learning success has been compared by [132], who identified the most important ones in both flipped courses and MOOCs. Empirical evidence of their importance across MOOCs was also provided by [151]. In our study, we built our models based on the indicators proven to be important in previous work, specifically focusing on the dimensions of effort regulation, time management, and metacognition. The granularity and comprehensiveness of the collected log data allowed us to consider these dimensions. However, it is important to note that critical thinking and help-seeking could not be directly measured in our study. The learning indicators we considered can be categorized into the following dimensions:

Control: This dimension includes indicators related to in-video and cross-video behavior, which serve as proxies for a student's ability to control the cognitive load during video lectures and demonstrate metacognitive skills. Examples of indicators in this dimension include the proportion of videos watched, re-watched, or interrupted, reflecting the flow of learning and the student's ability to segment their learning process [138, 139].

Effort: The effort dimension focuses on monitoring the level and frequency of student engagement with the course content, including both videos and quizzes, as it has been proven to be fundamental for learning success [131, 134]. Indicators in this dimension include the total number of student clicks on weekends and weekdays, as well as the total number of study sessions.

Proactivity: This dimension aims to measure the extent to which students are proactive and stay on schedule, which has shown to be predictive of performance, especially in MOOCs [132]. The indicators in this dimension are related to the completion of videos and quizzes, based on the scheduled week of the course. Example features include the number of scheduled videos watched for a given week and the number of quizzes passed on the first try.

Regularity: The regularity dimension is associated with time management and captures patterns of student engagement within a week and throughout the day. These indicators

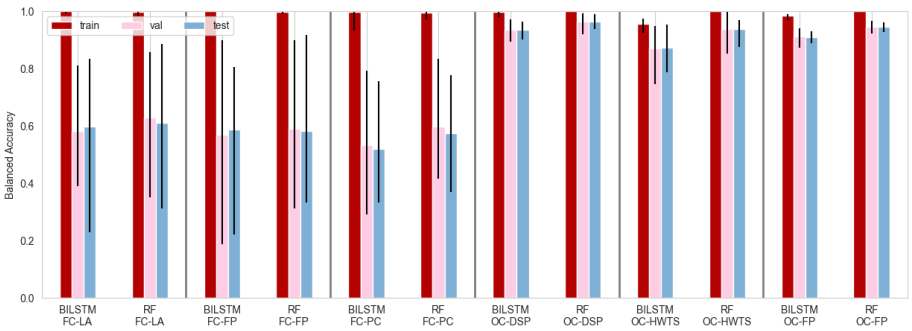


Fig. 5.3: Student Success Models Performance

reflect whether a student consistently engages on specific weekdays or at specific times of the day, which has been shown to be predictive of student success in MOOCs and flipped classrooms [135, 136]. Since the scores of the indicators can vary, we performed min-max normalization for each feature across all students and weeks. While the selected indicators fall into four dimensions and include several measures per dimension, we acknowledge that other relevant dimensions and measures may be beneficial for student success modeling. Our study can be easily extended to include a different (and larger) set of behavioral indicators.

Student Success Modelling - Step 3 / 6

A wide range of student success models have been proposed so far in the literature [153]. To align with prior work, since our study does not aim to propose a novel model, we considered two models reporting a high accuracy while providing a certain level of interpretability. Random Forest (RF) classifiers have achieved this in both flipped and MOOC contexts, when fed with behavioral features [154, 132]. Recent neural network classifiers based on BiLSTMs including attention layers, sigmoid activation, and a cross entropy loss function, have resulted in higher accuracy [151] and good interpretability as well [155]. Again, we based our decision on the similarity of the underlying context and logging system. We acknowledge that other models, e.g., Linear Regression and Support Vector Machines, have been used in prior works (e.g., [135]), but we left their analysis as a future work, using RFs as a representative of this class of models.

For each course and model, we applied a nested student-stratified 10-fold cross-validation. The same folds were used for all experiments across models, and hyper-parameters were optimized using grid search. In each iteration, an inner student-stratified 10-fold cross-validation was performed on the training set to select the combination of hyper-parameter values that yielded the highest accuracy on the inner cross-validation.

A total of 200 models per course were obtained (2 architectures × 10 outer folds × 10 inner folds). The balanced accuracy was computed on the training and validation sets to

assess model validity and on the test set to evaluate model generalizability. The averaged balanced accuracy on the three sets for each course and architecture combination is shown in Figure 5.3. The models performed well on both flipped courses and MOOCs, with slightly higher accuracy observed for models predicting on MOOCs. The error bars represent the average, maximum, and minimum balanced accuracy achieved by the random forest and bidirectional LSTM models through a nested 10-fold cross-validation for each course included in our study.

The models were able to predict the probability (p) that a student would fail the course. A decision threshold of 0.5 was used to obtain the final predicted label: $\tilde{y} = 0$ if $p < 0.50$ (predicted pass) and $\tilde{y} = 1$ if $p \geqslant 0.50$ (predicted failure). This threshold is commonly used in machine learning and prior education-related works.

The model confidence (c) was determined by measuring the proximity of the predicted probability (p) to the decision threshold. The model confidence value (c) is calculated as $|p - 0.50|$, ranging between 0 and 0.50. Higher values indicate a higher level of confidence in the prediction.

Prediction Student Grouping - Step 4 / 6

The practice of obtaining a trained student success model and computing importance scores for each indicator to link them to successful patterns has been commonly observed in the literature for both RFs [132] and linear regression models [135]. Recent work has also utilized explainability methods to extract these scores [155]. However, these approaches assume that the model accurately captures the relationships between learning and students' success, which may not be true for a significant portion of students. It is crucial to consider this portion of students to ensure that no student is negatively impacted by the model.

Figure 5.4 illustrates an emerging issue related to model confidence. Each plot shows the predicted probability distribution for students in the training, validation, and test sets, with 200 pass and 200 failing students randomly selected. In the case of flipped courses, BiLSTM models often exhibit high confidence in their predictions, as indicated by the skewed distribution towards the two extremes. However, their balanced accuracy on flipped courses is not high, as shown in Figure 5.3. In the case of MOOCs, the same models demonstrate high confidence, and their accuracy is also high. Interestingly, RF models on flipped courses exhibit high uncertainty in predictions for unseen validation and test students. These results raise concerns about the possibility of the model being highly confident but incorrect, which can have serious consequences when models are used in real-world applications and influence human understanding of the learning process.

Given these observations, it becomes essential to investigate the relationship between the correctness of predicted labels and the model confidence. Let us assume a hypothetical trust level $\delta \in (0, 0.50)$ that users have while using model predictions. Predictions

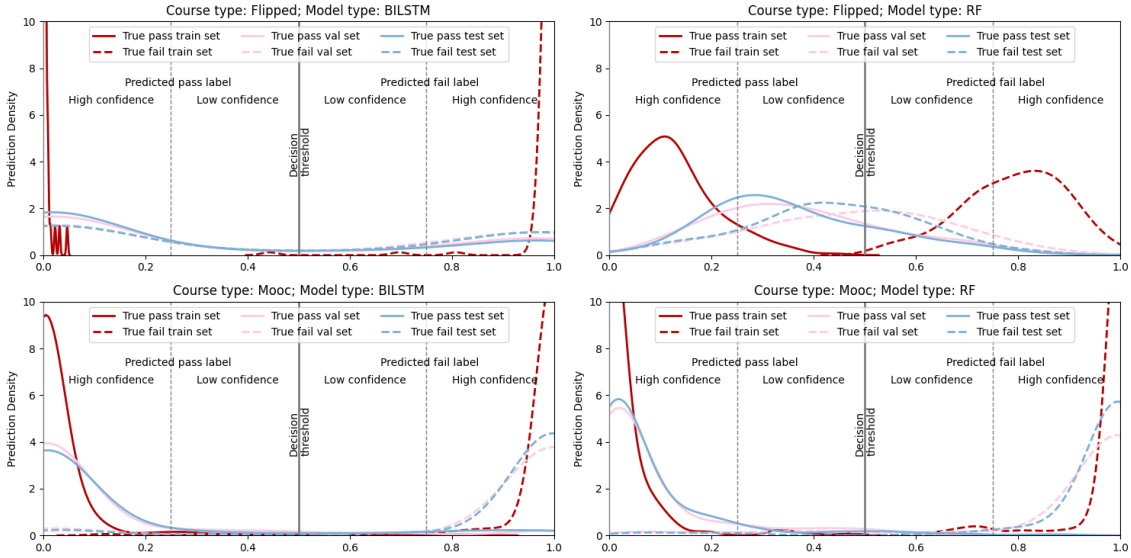


Fig. 5.4: Student Success Predicted Probabilities.

with a confidence $c \geqslant \delta$ would be considered more trustworthy, while predictions with $c < \delta$ would be considered less trustworthy. Based on the true label y and the predicted label $\tilde{y}$ with confidence c for a given student, the student can be assigned to one of the following groups:

- **Known Knowns** ($g = 0$): Students for whom the model is highly confident ($c \geqslant \delta$) and the predictions ($\tilde{y}$) correctly reflect the true success or failure ($\tilde{y} = y$). These are ideal cases where the model is correctly optimistic, providing high-confidence predictions that align with the true outcome.

- **Known Unknowns** ($g = 1$): Students for whom the model is not sure ($c < \delta$) and any predicted label ($\tilde{y}$). These cases require extra caution before using the predictions to take actions in the real world. The concept of known unknowns accounts for errors expected based on the probability estimates of the classification.

- **Unknown Unknowns** ($g = 2$): Students for whom the model is confident ($c \geqslant \delta$) but the predictions ($\tilde{y}$) are actually incorrect ($\tilde{y} \neq y$). Intuitively, these examples are far from the decision boundary but have been labeled incorrectly. Although these examples may be rare, their non-negligible prevalence poses a risk.

For convenience, we used a trust level $\delta = 0.25$ in our study, and analysis under other trust levels is left as future work. It is important to note that this threshold can be adjusted according to the model's accuracy and the requirements of the final system. In our study, we considered the same threshold for both models to ensure a fair comparison.

Prediction Group Characterization - Step 5 / 6

To address unknown unknowns, one might consider classifying future students into the category of unknown unknowns. This could involve training a second model to predict whether a student belongs to this category and taking appropriate action based on the prediction. However, this solution has several weaknesses, including increased complexity, potential biases, and limited data availability. Instead, it is more effective to make stakeholders aware of unknown unknown cases and provide them with information about typical archetypes of unknown unknowns. Instructors can then use this information to improve data collection and processing. The choice of variables used to explain unknown unknown cases is a crucial aspect in this scenario.

There are two complementary approaches to selecting these variables [156]. In confirmatory research, the potential impact of different variables is hypothesized based on existing theories. In an exploration-driven approach, variables are selected when there is a lack of theories or when generalizing across domains. Given the recent awareness of unknown unknowns, exploration research can generate new hypotheses to be later evaluated by experts, such as instructors, in their courses.

To characterize unknown unknowns from the model's perspective, we examined the relationship between (i) model confidence and predicted label correctness based on the student groups identified in the previous step, and (ii) a range of variables known to the model, namely the behavioral indicators used for training. To provide context, we also considered certain demographic variables that were not included in the model training for ethical reasons but may be available as contextual variables. Formally, the independent variables for our characterization were represented by a vector v for each student, including the averaged values of the 45 behavioral indicators across course weeks, as well as the gender and provenience attributes. The dependent variable was the group label g computed in the previous section.

Our goal was to identify variables that have a statistically significant relationship with the dependent variable. We employed multiple regression analysis, fitting the model with the input vectors v and the corresponding group labels g. Formally, the equation was $y = \epsilon + \gamma_0 + \sum_j \gamma_j \cdot v_j$, where y represents the dependent variable, γ_j denotes the regression coefficient, and v_j represents the value of the j-th variable. The coefficients $\gamma \in \Gamma$ provide insights into the extent to which changes in a given variable, while holding all others constant, lead to changes in the group membership. By conducting the multiple regression analysis, we tested the null hypothesis that all coefficients $\gamma \in \Gamma$ are zero, against the alternative hypothesis that at least one coefficient γ_j is nonzero. We hypothesize that variables with nonzero statistically significant coefficients can be considered important for explaining membership in the unknown unknown group.

Table 5.2: Content of the questionnaire provided to instructors.

ID	Question on Demographics
Q01	Which type of organization are you based in?
Q02	Which role do you have in your organization?
Q03 (Q04)	Which continent (country) are you based in?
Q05	Which age group are you in?
Q06	With which gender identity do you most identify?
Q07	How many courses were you teaching assistant for?
Q08	How many courses were you an instructor for?

ID	Question on Student Success Models
Q09.1 (Q10.1)	Be (less) confident about using model predictions.
Q09.2 (Q10.2)	Feel assisted (more) in diagnosing students' difficulties.
Q09.3 (Q10.3)	Trust (less) using model predictions in your classroom.
Q09.4 (Q10.4)	Feel (more) successful in using model predictions.
Q09.5 (Q10.5)	See as (less) transparent how model predictions are made.
Q09.6 (Q10.6)	Rely on model predictions at least as much as you rely on a recommendation from a colleague.

Instructors' Opinion Collection - Step 6 / 6

Finally, we aimed to understand how instructors' perception of student success models would change when they become aware of unknown unknowns. To achieve this, we designed a questionnaire [1] divided into two main sections (see Table 5.2), following prior work on trust in artificial intelligence for education [146], including two main sections (see Table 5.2): one on demographic information, whereas the other was focused on a use case on unknown unknowns in student success modelling.

In the first part, we were interested in knowing who the experts were, including the type of their organization (e.g., academia, industry), their role in the organization (e.g., full professor, researcher), the continent and country they are based in, the age group (in a specific range), and the gender identity. In addition, we asked participants how many courses they acted as an assistant (e.g., tutor) and as an instructor (e.g., full professor, associate professor) in.

In the second part, we described a use case with student success predictions presented to the instructor while investigating which students would require assistance. This use case was accompanied by a provocative user interface which would require them to think about the influence of student success models on the instructor[2]. First, we asked experts to rate their perceived confidence, assistance, success, trust, transparency, humanity

[1] The full questionnaire is available at the following webpage: https://shorturl.at/qvX47.

[2] The user interface was on purpose limited to a rudimentary pass/fail setting, without any confidence level, to stimulate instructors' critical thinking and reduce the impact of other user interface elements on the unknown unknowns perception.

with respect to this case. This decision let us understand the original perception of instructors towards student success models in general to better contextualize our results. We then explained the concept of unknown unknowns and envisioned a simple addition to the user interface, indicating students at risk of being unknown unknowns via alerts and explanations. This information would trigger instructors to question the provided prediction and investigate the corresponding students' behavior more thoroughly. We then asked the extent to which their perception of the student success model changed relatively to their original perception (e.g., more or less trust). In both series of questions, each participant was allowed to select among four possible answers, from "strongly disagree" to "strongly agree". No neutral question was introduced; hence, answers should be interpreted as what the instructors would answer in case they were forced to make a decision. As a final field, we asked if they identified any (dis)advantages raised by being aware of unknown unknowns.

Following [10]'s protocol, we e-mailed the questionnaire to experts with a paper accepted in a top conference in education in 2021 (AIED, EAAI, EC-TEL, EDM, ICALT, ITS, LAK, L@S). Out of 1,721 people, 112 (6,51%) completed the questionnaire. This choice was made to include people who are both educators and experts in the field. Though we acknowledge that future work will need to focus also on the perception of a generic instructor, unknown unknowns analysis is still at early stages and our feedback can make more mature our understanding before involving them.

5.1.3 Experimental Results

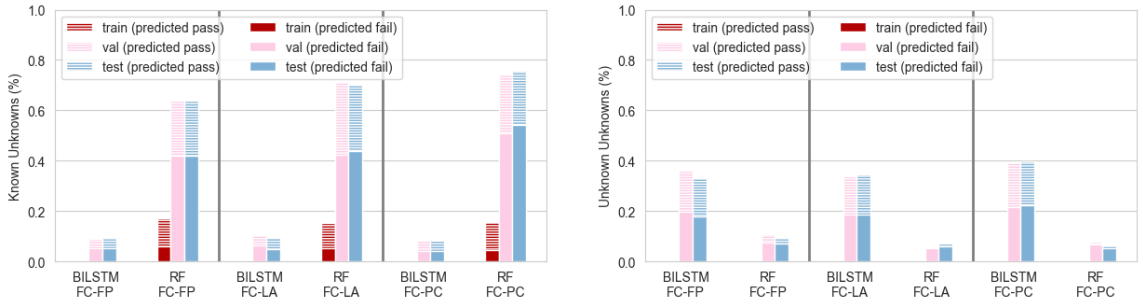
Although our methodology can be used to analyze several perspectives, we focused on the following research questions:

- **RQ1:** Do unknown unknowns exist in success prediction? How do they vary across flipped courses and MOOCs?
- **RQ2:** Can we characterize common unknown unknowns in the considered student success prediction models?
- **RQ3:** How does providing instructors with signals about unknown unknowns impact their perception of the model?

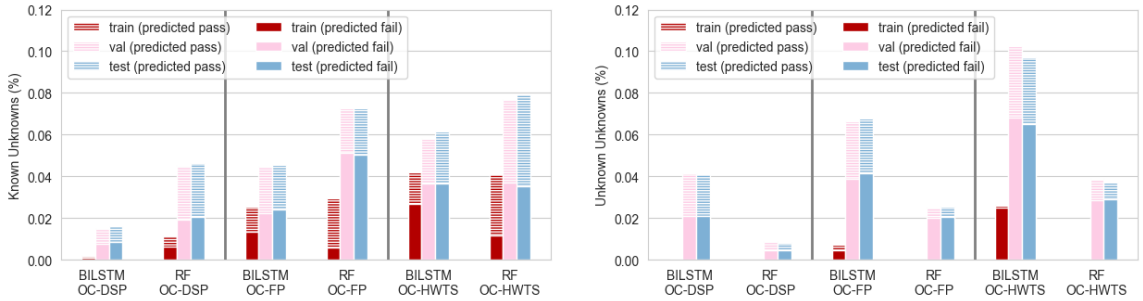
By investigating these questions, we sought to gain insights into the presence and characteristics of unknown unknowns in student success prediction models, as well as the potential influence of providing instructors with information about these unknown unknowns.

RQ1: (Un)known Unknown Prevalence

In our first analysis, we aimed to investigate the existence of unknown unknowns in the student success models we developed. We also wanted to explore whether there



(a) Flipped courses



(b) MOOCs

Fig. 5.5: Prediction Student Groups: AVG percentage of students in train, val and test.

were any differences in the prevalence of unknown unknowns between courses of the same type and across different course types.

Figure 5.5 provides insights into the average percentage of students belonging to the known unknown and unknown unknown groups in the in the training, validation, and test sets being a known unknown (left) and unknown unknown (right). Solid (dashed) bars indicate students who passed (failed) but were predicted as failing (passing). As expected, based on the model performance depicted in Figure 5.3, we observed a higher presence of unknown unknowns in flipped courses compared to MOOCs. To gain a better understanding, we analyzed the patterns in detail for each course type.

For flipped courses, we found that RF and BiLSTM models exhibited different behaviors regarding known unknowns and unknown unknowns. RF models were more conservative, resulting in a higher percentage of known unknowns. This suggests that RF models require instructors to consciously inspect students for whom the model is not confident in its predictions. On the other hand, BiLSTM models displayed higher confidence overall, but this confidence often led to incorrect predictions, resulting in a higher percentage of unknown unknowns. In flipped courses, the distribution of unknown unknowns was roughly equal between false failing and false passing predictions. These patterns were consistent across the flipped courses, indicating that the model's impact was more significant than the specific characteristics of the course itself.

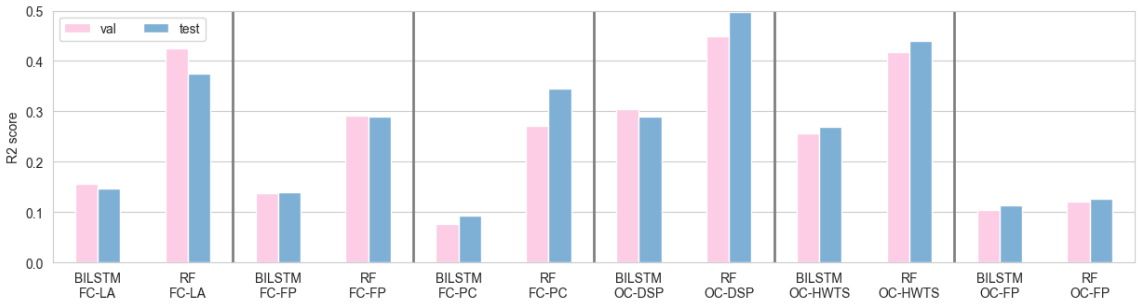
Similar patterns were observed in students attending MOOCs, although it is worth noting that MOOCs typically had larger student populations. For instance, in the case of the OC-DSP MOOC, a 2% prevalence of unknown unknowns would involve over 300 students. Similarly, a 4% prevalence in the OC-HWTS MOOC would affect more than 100 students.

Findings RQ1. *Unknown unknowns non-negligible prevalence was observed in both course types. Flipped courses were more prone to unknown unknowns than MOOCs. Given their comparable accuracy, RF models led to less unknown risks than BiLSTM models. Furthermore, the distribution of unknown risks was similar between false failing and false passing predictions.*

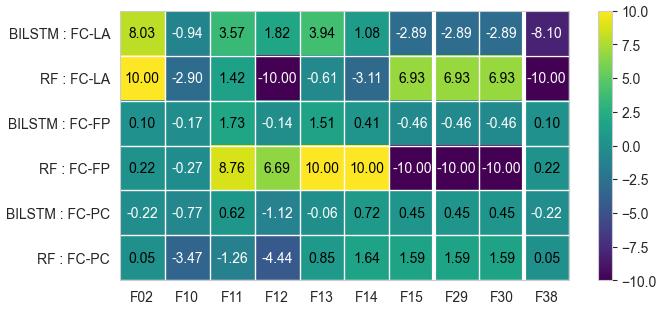
RQ2: Unknown Unknowns Characterization

In our second analysis, we aimed to characterize the unknown unknowns in the student success prediction models. We employed the explanatory framework for each combination of models and courses.

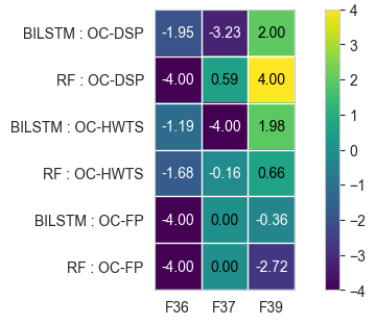
Figure 5.6a presents the average R2 scores, which indicate the proportion of variance in the dependent variable (unknown unknown membership) that can be explained by the independent variables (behavioral indicators). We observed that the variance was more



(a) R2 scores



(b) Highest coefficients on flipped courses



(c) Highest coefficients on MOOCs

Fig. 5.6: R2 score of the linear regression models and AVG coefficients

predictable in RF models than in BiLSTM models, and MOOCs had a higher predictability compared to flipped courses. Interestingly, none of the models achieved a reasonably high R2 score for the OC – FP course, indicating that the variance in unknown unknown membership was not well explained by the selected indicators. This finding confirmed our hypothesis that training a supplementary model specifically for unknown unknowns detection would not be a viable solution.

To further analyze the behavioral indicators important for predicting unknown unknowns, we examined the regression coefficients. Figures 5.6b and 5.6c summarize the indicators with the highest coefficients for each course type. For conciseness, we focused on indicators whose coefficients had an average value higher than 1 across courses of the same type. See Table 5.3 for a description of the indicators.

For flipped courses, we identified ten important behavioral indicators for explaining unknown unknowns. The indicators primarily belonged to the control dimension, with indicators related to engagement and proactivity playing a less prominent role. For example, indicators such as the frequency of play (F10), stop (F11), and speed events (F15), as well as the total number of video (F29) and problem clicks (F30), exhibited a high positive weight for being classified as unknown unknowns. We observed that the majority of students who passed the course tended to perform these actions frequently. The model struggled to determine whether these actions were indicative of students' struggles or their engagement and reflection on the content. As a result, the model often classified struggling students as passing with high confidence. Other indicators, such as the frequency of stop events (F12) and alignment with the schedule (F38), showed a very low negative weight. Being unaligned with the schedule did not necessarily imply that a student would fail the course. For instance, a student might have consistently learned offline, which the model was unaware of, leading to unknown unknowns.

In the case of MOOCs, our explanatory analysis revealed three relevant indicators, all related to proactivity. Notably, there was no overlap in the selected regression coefficients between flipped courses and MOOCs. Content anticipation, in particular, showed a strong association with a high regression coefficient. For students who were eager to learn, anticipating course content was seen as a positive attitude toward passing. However, for other students who were merely interested in previewing future content, interacting with the content in advance of the schedule might not be indicative of success. Additional information would be required for the model to distinguish between these cases.

Comparing the two model types within the same course, we found that different behavioral indicators influenced the models' predictions in the FC – LA course. While the BiLSTM and RF models agreed on a few indicators, such as F02, F10, F11, and F38, they strongly disagreed on the remaining six indicators. The two models showed more similar weightings for the behavioral indicators in the other two flipped courses. Similar patterns were observed for the regression coefficients in MOOCs. Interestingly, none of the considered demographic attributes were found to be important, highlighting the

Table 5.3: Behavioral Indicators of Unknown Unknowns.

Dimension	Indicator	ID	Description
Control	AvgWatchedWeeklyProp	F02	The ratio of videos watched over the number of videos available.
	FrequencyEventPlay	F10	The frequency between every Video.Play action and the following action.
	FrequencyEventPause	F11	The frequency between every Video.Pause action and the following action.
	FrequencyEventStop	F12	The frequency between every Video.Stop action and the following action.
	FrequencyEventSeekBackward	F13	The frequency between every Video.SeekBackward action and the following action.
	FrequencyEventSeekForward	F14	The frequency between every Video.SeekForward action and the following action.
	FrequencyEventSpeedChange	F15	The frequency between every Video.SpeedChange action and the following action.
Engagement	TotalClicksProblem	F29	The number of clicks that a student has made on problems this week.
	TotalClicksVideo	F30	The number of clicks that a student has made on videos this week.
Proactivity	CompetencyAlignment	F36	The number of problems this week that the student has passed.
	CompetencyAnticipation	F37	The extent to which the student approaches a quiz provided in subsequent weeks.
	ContentAlignment	F38	The number of videos this week that have been watched by the student.
	ContentAnticipation	F39	The number of videos covered by the student from those that are in subsequent weeks.

need for additional contextual variables that are often not recorded in the log data or the university's database.

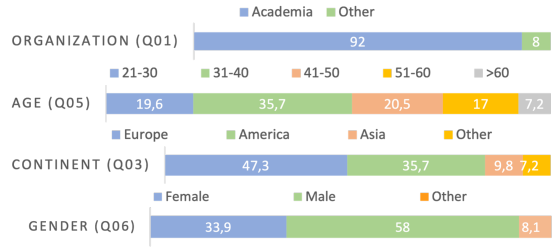
Findings RQ2. *Unknown unknowns membership was connected with certain behavioral indicators. Control (more), engagement and proactivity (less) characterized them in flipped courses. Concerning MOOCs, proactivity was shown to be the main dimension indicative of the models' unknown unknowns, regardless of the MOOC.*

RQ3: Unknown Unknowns Perception by Instructors

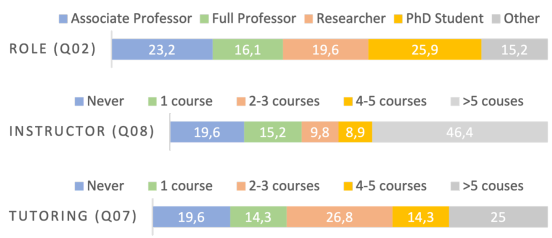
In our final analysis, we aimed to understand the instructors' perception of student success models when they were made aware of unknown unknowns. We distributed the questionnaire and collected responses from experts worldwide. Figure 5.7 summarizes the results obtained from the questionnaire.

Figure 5.7a provides an overview of the demographic distribution of our sample. The majority of participants worked in an academic context (92%), with a relatively balanced distribution among different roles, including associate professors, PhD students, and researchers. Our sample also exhibited a balanced distribution across age groups, with a slight predominance of individuals in the 31-40 age group. In terms of gender identity, the responses were predominantly from men (58%). Regarding teaching experience (Figure 5.7b), we found that 19.6% of participants had no experience as tutors or instructors, while 46.4% had served in more than five courses as a leading instructor or teaching assistant. The remaining participants reported teaching experience ranging from two to five courses. Overall, our sample displayed a good level of diversity in terms of demographics and teaching experience.

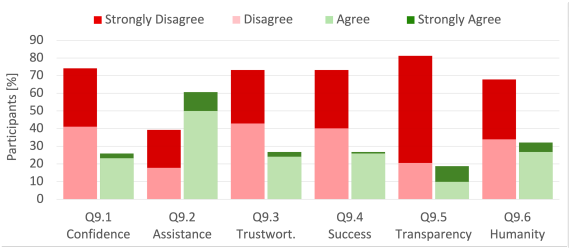
In the second section of the questionnaire, we focused on the instructors' perception of an example student success model (Q09). Figure 5.7c shows the results, indicating that a high percentage of participants expressed concerns about the model. The major-



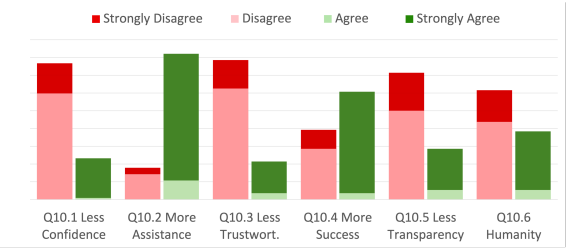
(a) Demographic distribution



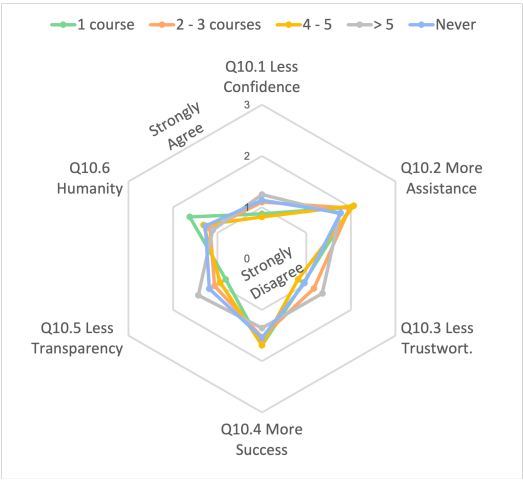
(b) Teaching experience distribution



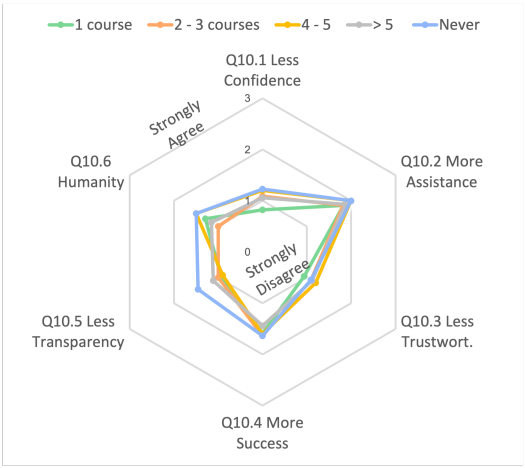
(c) Initial instructors' perception



(d) Change in Perception with UU Awareness



(e) Tutoring-wise relative change



(f) Instructor-wise relative change

Fig. 5.7: Instructors' Perception.

ity strongly disagreed or disagreed with statements related to confidence in the model predictions (74%), trust in using the predictions in the classroom (73.2%), perceived success in using the predictions (73.2%), transparency of the model predictions (81.25%), and reliability compared to a recommendation from a human colleague (67.86%). Notably, instructors felt more assisted by the model, as indicated by the lower percentage of participants who disagreed with the statement. It is important to note that our goal was to collect a baseline perception and not to assess the instructors' perception of student success models in general. This baseline perception served as a reference point for comparing the impact of unknown unknowns awareness.

In Figure 5.7d, we present the instructors' perception after they were provided with

signals and explanations about unknown unknowns in the user interface. To facilitate comparison, we asked participants to indicate the relative change in their perception across the six perspectives by alternating between positive (more) and negative (less) statements. The results showed that a high percentage of participants felt more confident (76.79% strongly disagreed or disagreed with the negative statement). They mentioned that the additional signals and explanations helped them understand the reasoning behind the predictions and increased their confidence in the model. Participants also felt more assisted (82.14% strongly agreed or agreed with the positive statement) and mentioned that the information provided allowed them to gain more insight into the model's predictions. Trust in the model's predictions improved as well, with 78.57% of participants strongly disagreeing or disagreeing with the negative statement. Instructors also reported feeling more successful in using the model predictions (60.71% strongly agreed or agreed with the positive statement). The perception of transparency improved, as 71.43% of participants strongly disagreed or disagreed with the negative statement. However, instructors still did not fully rely on the model predictions (61.61%) compared to human recommendations. Participants emphasized that while the additional alerts and explanations added valuable context, they would not completely rely on the model predictions but rather use them as an aid to identify students requiring attention. To provide a more detailed analysis, we cross-referenced these values with different demographic elements in Figures 5.7e and 5.7f. Interestingly, the findings showed that the additional alerts and explanations provided by unknown unknowns information improved instructors' perception by adding context and helping them assess and study students' learning behavior. Participants expressed that this would enhance the usability of the model.

Findings RQ3. *Being made aware of unknown unknowns information, instructors feel a higher confidence, assistance, trust, success, transparency. Though their perception slightly improved, instructors still did not truly rely on model predictions as much as human recommendation.*

5.1.4 Findings and Recommendations

With our experiments, we showed that unknown unknowns exist and vary in number and type across courses (**RQ1**). We then characterized unknown unknowns under the specific use cases (**RQ2**). Finally, we found that making instructors aware of unknown unknowns had a positive impact (**RQ3**). Our findings led to multiple implications.

Scientific Implications. The collected data is usually partial and does not provide the global picture of students' behavior, skills, and needs. Many other variables can be hardly collected for being included in the model reasoning. For instance, a student might work a lot offline and still get very good exam grades. Being the predictions dependent on the data, there is the high risk that at the moment no enough data for the model is collected. Our study therefore calls for a more extensive data collection aimed to bridge the gap

between what the models knows and what should know.

Our study has shown that, BiLSTM models were often very confident about their predictions, but completely wrong in several cases. When possible, model uncertainty would be preferred to avoid misleading instructors, as in RF models. Understanding what is the source of such model confusion and how this knowledge can be induced into the model is urging. Another implication is therefore the need of student success models that reduce unknown unknown cases.

Notably, we proved that assessing model performance solely based on accuracy may introduce unknown risks. When a student success model is delivered, evidence on its unknown unknowns should be provided. This evidence could be both quantitative, by reporting for instance the percentage of unseen students resulting as (un)known unknowns, but also qualitative, by characterizing the cases where the model is less confident but incorrect.

Instructors will be likely to use success models as a complementary support to their personal perception. Human-in-the-loop approaches can be used to let the instructor and the model help each other while identifying students that require assistance. Unknown unknowns represent examples the instructor should reflect on while using predictions.

Technological Implications. Our study has proven that confidence levels are not enough to prevent undesired behavior like unknown unknowns. Signals of their presence, though helpful, would just be triggers for further analysis of certain students. Indeed, pass/fail predictions alone would not be enough and more insights about learner behavior will be needed. Behavioral patterns (e.g., late submissions) can then indicate how to counsel students.

In our work, we adopted BiLSTM models, with sigmoid and cross entropy. that could tend to push predictions towards the two sides, 0 and 1. Our findings show that this practice creates more unknown unknowns (more risks). Future work should carefully consider this aspect while selecting model parameters to reduce unknown risks.

Once unknown unknowns patterns are identified, it will be important to understand how information about them should be presented. Having signals and explanations has led in our study to an increase of trust in the model prediction. Rather than a signal, showing a null state when a prediction is potentially risky would be another solution. Instructors might feel that the model did not have enough data to learn the likely outcome for some of the students yet.

Besides being used for learning understanding, student success models can fuel tools to recommend instructors student requiring assistance (especially in large classes). It could be also important to have grouped predictions, along with individual predictions, based on similarities. Reducing unknown risks will be even more important in these settings.

Social Implications. In our results, we showed that raising awareness of unknown unknowns led to higher confidence, feeling of assistance, trust, and transparency. Being trustworthiness a complex challenge in this field, our study introduces another source for improvement, complementary for instance to explanations. Furthermore, such models could be harmful by creating false expectations. Being aware of unknown unknowns can help to prevent such situations.

Finally, predicting whether a student is going to pass or fail a course is of practical utility if it is done early. Once all of the data is collected over the course end, the models' purpose might be just to detect at-risk students who might benefit from remedial sessions between the course end and the final exam. Nevertheless, we believe that the unknown unknowns might be even more evident in very early predictions, which we plan to investigate in future work.

Scientific, technological, and social shifts in education often go hand-in-hand. Countering unknown unknown issues will be essential to further strengthen the reliability of emerging student success models in real-world education.

5.2 Model-Human Comparison on Student Success Prediction

5.2.1 Introduction

Educational institutions have embraced various modes of instruction that combine or replace face-to-face lectures with online activities. Examples include flipped learning [142] and distance learning [143], which differ in the sequencing of face-to-face and online sessions. However, when education partially or fully transitions online, teachers face challenges in monitoring and adapting their teaching practices to ensure that every learner receives adequate support [157, 158].

To support teachers in identifying and assisting learners in need, strategies utilizing learning analytics methods have gained prominence [133]. One common approach is the use of visualizations and dashboards to provide insights into the learning process [159]. In more advanced cases, automated models can predict whether learners are at risk of not achieving their expected learning goals [160]. Student success prediction models, in particular, have emerged as a valuable tool for creating personalized learning experiences by modeling student performance.

However, the process of identifying learners in need of support is complex for both teachers and models. Visualizations can sometimes overwhelm teachers, making it challenging for them to determine where to focus their attention and what is most important [161]. Models, on the other hand, can develop weaknesses due to unknown information, unmodeled biases in the data, shifts in data distribution, or suboptimal algorithmic

choices [162, 10], resulting in low trust from teachers. This lack of knowledge can lead both teachers and models to be confident in their predictions but ultimately incorrect [13]. While efforts have been made to explore unknown unknowns in other domains, such as medical diagnosis [163], workflows [150], and financial services [164], to the best of our knowledge, we are the first to uncover unknown unknowns issues in education [13]. Unfortunately, no research study exists that addresses the gap between teacher and model predictions, hindering the development of strategies to mitigate the impact of unknown unknowns in student modeling.

In this work, we investigate the dissonance between teacher and model predictions regarding learners in need, based on their behavior [14]. Teachers, being human, may be better equipped to handle abstract and subjective tasks, while models struggle with such nuances. We aim to examine whether teachers consistently outperform models in making decisions related to educational tasks, specifically the prediction of at-risk learners in a flipped course based on their behavior in pre-class sessions. Our work is guided by three research questions:

- **RQ1:** How do teachers, compared to a well-known model, determine whether a learner is at risk?
- **RQ2:** Is there a relationship between the confidence of teachers/models and the correctness of their decisions?
- **RQ3:** What additional knowledge and intervention needs do teachers identify?

To address these research questions, we employ a crowdsourcing approach with 360 human intelligence tasks from 60 university teachers. We provide teachers with visualizations related to various dimensions of self-regulating behavior for each learner and ask them to predict the probability of the learner failing the course. We also gather information about the rationale behind their decisions, their level of confidence, and any additional knowledge or intervention needs they may have. By comparing teacher and model decisions in terms of correctness and confidence, and analyzing the teachers' rationale and needs, we gain insights into the dissonance between human and model decision-making processes. Our results provide valuable insights on how to responsibly use predictions from student performance models to build personalized student models based on performance, highlighting (dis)agreements between machine-learning models and teachers as well as weaknesses in their accuracy.

5.2.2 Methodology

In this section, we describe the educational context and the approach we adopted to gather both model and teacher predictions and their reasoning. Figure 5.8 provides an overview of all the steps of our methodology.

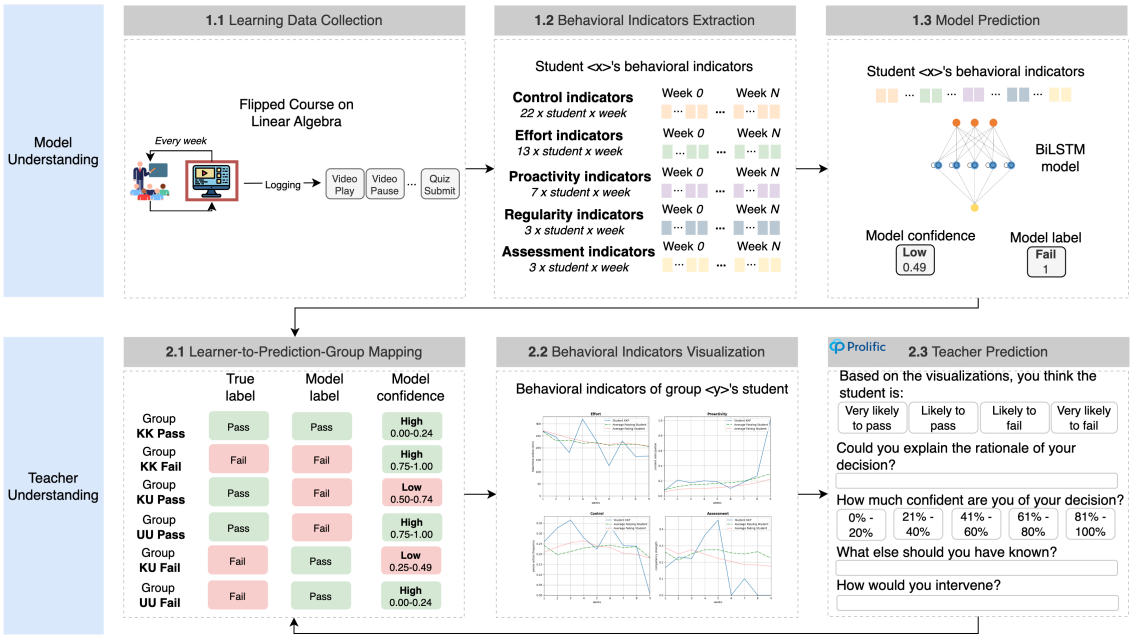


Fig. 5.8: Methodology steps for human understanding.

Model Understanding

In a first stage, we describe the machine learning pipeline adopted to gather the model understanding, including the learning data collection, the behavioral indicators extraction, and the model creation and prediction.

Learning Data Collection. Our study focuses on a Linear Algebra course delivered in a flipped format in a European university to 292 learners. Specifically, we focus our analysis on a semester-long Bachelor course with a 14 weeks schedule composed of lectures and sessions of recitation or exercises [152]. In-class activities included short problem-solving exercises and structured proof-type problems. In addition, learners were expected to spend some hours per week on individual study (a list of sections from a MOOC with video lectures and online quizzes) as a preparation for class (pre-class activities). Instructions on preparatory work were sent to learners a week in advance. In total, 179 quizzes, normally multiple-choice questions, allowed learners to self-evaluate their own learning at home. The MOOC platform collected data on learners' pre-class activities. Log entries reported the user (e.g., user 0), activity (e.g., user 0 play video x), and timestamp (02-05-2022 11:03:00). In total, we considered more than 145,640 log entries. There was no data recorded on in-class activities. Learner performance was measured according to the exam grade (failing rate 40%). The study was approved by the university's ethics committee (HREC No. 058-2020/10.09.2020).

Behavioral Indicators Extraction Prior work has found significant association with academic achievement for self-regulated learning (SRL) aspects [63, 131, 165], including

effort regulation (persistence in learning), time management (ability to plan study time), metacognition (awareness and control of thoughts), critical thinking (ability to carefully examine material), and help-seeking (obtaining assistance if needed). A variety of learning indicators' set have been proposed in the literature accordingly. Their power for modeling learning success has been compared by [166]. This work identified the most important ones in both flipped courses and MOOCs. In our study, we followed a similar logging policy and experimental scenario as the aforementioned work. Therefore, we built our models based on the indicators that were found to be important in their study. The granularity and comprehensiveness of the collected log data allowed the studies mentioned above, as well as ours, to consider effort regulation (effort), time management (regularity and proactivity), and metacognition (control, assessment) dimensions. Specifically, we considered the following indicators.

The indicator *control* (22 indicators per learner per week) models in- and cross-video behavior as a proxy of learner ability to control cognitive load through weeks (metacognition). For instance, in-video flow, manageable through the platform functionalities (e.g., pause button), could include regular pauses to segment learning [138, 139]. Among others, this feature's set consists of the proportion of videos watched, re-watched, or interrupted. *Effort* (13 indicators per learner per week) aims at monitoring how much and how frequently learners engage with the course content (both videos and quizzes) and is proven to be fundamental for learning success [131, 134]. These features included indicators such as the total number of learner clicks on weekends and on weekdays, and the total number of sessions. *Proactivity* (7 indicators per learner per week) attempts to measure the extent to which learners are on time or ahead of the schedule and is demonstrated to predict performance especially in MOOCs [166]. These features are related to completion of videos and quizzes, according to the week of the course they are schedule in. Example features included the number of scheduled videos watched for that week and the number of quizzes passed on the first try. *Regularity* (3 indicators per learner per week) is also associated with time management. It estimates the intra-week and intra-day time management patterns (i.e., capturing whether a learner regularly engages on specific weekdays or day times), proven to be predictive of learner success [135]. Finally, the assessment dimension (3 indicators per learner per week) assumes that there is a relation between learner performance in voluntary non-graded online quizzes and the final course grade (e.g., [166]).

Since indicator scores vary in their range, we performed a min-max normalization per feature across all learners and weeks for that feature. While the selected indicators cover five dimensions and include multiple measures within each dimension, we acknowledge that there may be other relevant dimensions and measures that could be beneficial for modeling learner success. The choice of indicators was based on prior research and their proven significance in predicting learning outcomes. However, it is important to continuously explore and evaluate additional dimensions and measures that could enhance the accuracy and effectiveness of success modeling in the educational context.

Model Prediction A wide range of learner success models have been proposed so far in the literature [66]. Since our study does not aim to offer a novel model, to align with prior work, we considered a model reporting a good accuracy while providing a certain level of interpretability. To predict learner success, we utilized a neural network classifier based on BiLSTMs with attention layers, sigmoid activation, and a cross entropy loss function. This choice was motivated by the model's good accuracy and interpretability, as demonstrated in prior work [167, 155]. Additionally, we considered the similarity of the educational context and logging system to ensure relevance and comparability with existing studies. We acknowledge that other models have been used in prior works (e.g., [135]), but we postponed their analysis as a future work. In order to train models, we applied a nested learner-stratified (i.e., dividing the folds by learners) 10-fold cross validation. In each iteration, we ran an inner learner-stratified 10-fold cross-validation on the training set, and selected the combination of hyper-parameter values yielding the highest accuracy on the inner cross-validation. We then evaluated it on learners in the test fold to show reproducibility in the same context. Balanced accuracy varied between 68% and 75%, depending on the fold, with an average of 73%.

Teacher Understanding

To understand the teacher understanding of learner success and compare it to the model's predictions, we employed a crowd sourcing approach that involved the following main steps: learner sampling, creation of visualizations, and delivery of the questionnaire.

Learner-to-Prediction-Group Mapping To gather the teacher understanding, the first aspect to consider is related to which learners should be paired to the teacher. In our work, we decided to group learners based on the correctness and confidence of the model's predictions and then uniformly sample them from the corresponding groups. This design choice allowed us to offer teachers examples where the model performance varies, thus we have investigated whether the teacher performance is (dis)similar. Let us consider a learner known to have a true label y and a model that predicts a label $\tilde{y}$ with a confidence c. With a decision threshold of 0.50, values of $c \in [0.25, 0.50)$ and $c \in [0.50, 0.75)$ would be considered as low confidence for passing and failing, respectively. Conversely, values of $c \in [0.00, 0.25)$ and $c \in [0.75, 1.00)$ would be considered as high confidence for passing and failing, respectively.

Based on the assumptions above, we assigned learners to one of the six groups in Table 5.4. In the first row, we reported two groups of learners for whom the model was confident, either $c \in [0.00, 0.25)$ or $c \in [0.75, 1.00)$, and the prediction correctly reflected their true success or failure. These were the ideal cases where the model was correctly and confidently optimistic, i.e., the model reported high confidence and correct label for the examples in this category. In the general machine learning field, examples belonging to this category are usually referred to as Known Knowns (KK). As a convention for this work, learners who actually passed the course ($\tilde{y} = y =$ Pass) are referred to as KK

Table 5.4: Model Prediction Groups.

True passing learners				**True failing learners**			
group	y	$\tilde{y}$	c	*group*	y	$\tilde{y}$	c
KK `Pass`	Pass	Pass	High [0.00,0.25)	KK `Fail`	Fail	Fail	High [0.75,1.00)
KU `Pass`	Pass	Fail	Low [0.50,0.75)	KU `Fail`	Fail	Pass	Low [0.25,0.50)
UU `Pass`	Pass	Fail	High [0.75,1.00)	UU `Fail`	Fail	Pass	High [0.00,0.25)

`Pass`. On the other hand, learners who actually failed the course ($\tilde{y} = y =$ Fail) are referred to as KK `Fail`. In the second row, we considered two groups of learners for whom the model was unsure, either $c \in [0.25, 0.50)$ or $c \in [0.50, 0.75)$, and the predictions would require extra care before being used. This concept mapped cases for which errors were expected based on the confidence of the model classification, being close to the decision threshold. In machine learning, examples belonging to this category are referred to as Known Unknowns (KU). Learners in this category who actually passed the course ($\tilde{y} = y =$ Pass) are referred to as KU `Pass`; whereas, learners who actually failed the course ($\tilde{y} = y =$ Fail) are referred to as KU `Fail`. Finally, the third row indicates two groups of learners for whom the model was confident, either $c \in [0.00, 0.25)$ or $c \in [0.75, 1.00)$, but its predictions are actually wrong. Intuitively, these are examples distant from the decision boundary yet labeled incorrectly. These examples, usually named Unknown Unknowns (UU), may be rare to be detected, but their non-negligible prevalence makes learners at risk. Learners in this category who actually passed the course ($\tilde{y} <> y =$ Pass) are referred to as UU `Pass`; whereas, learners who actually failed the course ($\tilde{y} <> y =$ Fail) are referred to as UU `Fail`. For details on this process, we refer the reader to [13].

Behavioral Indicators Visualization To communicate indicators to teachers regarding learners, we aimed for our study to resume a practical real-world situation. Machine-learning models have access to a large amount of data, to a potentially very large number of indicators, while teachers only have access to a subset of these indicators, due to well-known human characteristics about information processing and memory capabilities. This aspect has been highlighted in prior work in education in the context of learning dashboards [155]. To align our study with the real world, we fed the model with indicators relevant in prior work [168]. We based our selection on those that investigated such indicators' importance [63], concerning those to be shown to the teachers.

It was indeed impractical to show all of the over 40 indicators to each teacher in the form of visualizations. We instead opted for one representative of each SRL dimension (effort regulation, time management, metacognition, and assessment), found relevant in prior work [63], except for regularity[3]. Moreover, discarding regularity indicators was motivated by two reasons. First, prior work found that regularity is more predictive of success in MOOCs rather than flipped courses [166]. Second, indicators of regularity are

[3]We fully acknowledge that what and how we show information importantly influence the teacher reasoning and understanding. Therefore, we invite the reader to take this into account throughout the rest of the paper, especially while analyzing the results. Our findings open to future research on how the content of visualizations might influence both the correctness and confidence of teacher decisions.

usually hard to interpret by non-technical people, being based on complex mathematical functions such as Fourier transforms.

We selected four indicators among those provided to the model [166]. In terms of effort regulation, we included the total time spent online. This indicator was computed by summing the session durations in a given week, measured in minutes. For time management, we provided the teacher with an indicator pertaining to proactivity, i.e., content anticipation. It monitors the extent to which learners are on time or ahead of schedule, measuring it as the fraction of videos watched before the scheduled due date. The pause frequency was considered as the indicator of metacognition. Concretely, it was measured as the mean number of pauses, divided by the time spent watching a video, averaged across videos. Finally, for the assessment dimension, we considered the learners' competency strength, computed as the highest grade achieved on a quiz, divided by the number of attempts, averaged across quizzes.

Subsequently, we designed visualizations of the identified indicators according to the nature of the data as well as prior work on visual designs in education [159]. Being the indicators measured per week, they could be naturally represented as a time series. Hence, we adapted the most informative designs from [63]. To support reasoning, visualizations enabled teachers to compare the indicators of the current learner and those of the average passing and failing learner. Please refer to the survey at https://shorturl.at/bnGOU for an example student's visualizations presented to a teacher.

Teacher Prediction To assess teachers' understanding, we conducted a survey with 60 participants. We allowed individuals to participate in case they served as a lecturer in at least one university course. We recruited a balanced sample in terms of gender (45% identified as female) through Prolific, filtering those fluent in English. The most represented age group was 31-40 (41,67%). 68% of the participants came from Europe. They served as lecturers in one course (30%), two/three courses (33,33%), and more than five courses (28,33%). They were well-distributed across full professors (21,67%), associate professors (31,66%), and researchers (26,66%).

Our survey was organized into six sections, including 39 items (demographics: 5 items; data literacy: 4 items; visualization: 5 items x 6 students). In the first section, we explained the purpose of the survey, emphasized the target audience, and asked to agree with terms and conditions. Then, in the second section, we asked participants to provide demographic attributes (gender, age, country, teaching experience, role) to contextualize their answers. In the third section, participants were given the scenario of teaching a large university flipped course for over 300 Bachelor learners. They were told that they had taught the entire course and would be shown visualizations of their learners' SRL behavior to potentially identify intervention needs during the exam preparation period. In the fourth section, to assess participants' data literacy, we asked them to interpret visualizations of an example learner and answer to four understanding questions. In particular, for each visualization, we reported the definition of the corresponding indicator

and the interpretation of its value in a certain week. Based on this information, we asked the participant to answer to a simple question related to the plot, giving three possible answers ("yes", "no", or "I don't know"). The fifth key section included the visualization items grouped by learner, showing to the teacher one learner from each prediction group (six learners in total). For each student prediction group among the six described in Section 2.2.1 (see Table 1), we uniformly sampled four representative students for that group to be included in the student pool in the questionnaire, according to the model confidence in the respecting prediction. This choice allowed us to cover a more fine-grained level of cases that vary in the way the model behaves. While generating the questionnaire for a teacher, the student of a given group to be shown to the teacher was randomly sampled from the created student pool for that group. This choice was made to facilitate the questionnaire delivery and make the analysis easier to interpret. For each learner, teachers were asked to interpret the graphs by (1) making a passing or failing decision, (2) providing the rationale behind their decision, (3) their confidence in the decision, (4) any other information they would have known to refine their decision, and (5) any intervention they would have made in that case (the full survey is available at https://shorturl.at/bnGOU).

5.2.3 Experimental Results

To compare model and teacher decisions (**RQ1**), identify the relationships between confidence and correctness (**RQ2**), and emphasize any additional knowledge and intervention need (**RQ3**), we analyzed teacher answers to our survey.

RQ1: Teacher-Model Correctness Agreement

In a first analysis, we investigated how teachers compare to the considered model in determining whether a learner is at risk. We analyzed teacher predictions both at prediction group- and learner-level as well as their reasoning.

Group-level Correctness Fig. 5.9 collects the teacher decisions for each prediction group. We created two separate plots, one for the decisions on truly passing learners (Fig. 4a) and one for those on truly failing learners (Fig. 4b).

Fig. 4 collects the teachers' accuracy for each prediction group, created according to the considered machine-learning model performance. For convenience, we created two separate plots, one for the decisions on truly passing learners (Fig. 4a) and another one for those on truly failing learners (Fig. 4b). Each plot includes three bars representing teacher performance on the three respective prediction groups. Given a bar, the blue portion indicates the percentage of the teachers who predicted the student's outcome correctly, while the red portion refers to those who predicted the student's outcome incorrectly. To enable a more detailed understanding, for each blue / red portion of a bar, the solid sub-portion refers to teacher predictions that were unsure ("very likely"), whereas the dotted sub-portion refers to teacher predictions that were more sure

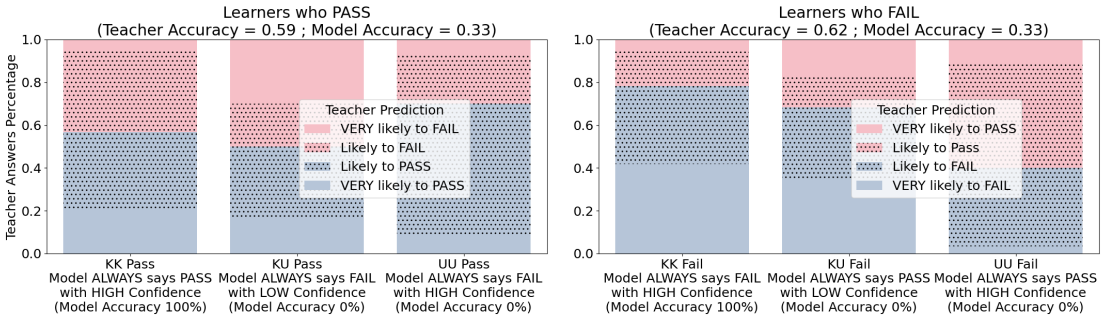


Fig. 5.9: Distribution of teacher prediction (RQ1)

("likely"). Teacher accuracy (blue portions) is calculated as the percentage of teachers providing the correct student outcome for each prediction group. As an example, the left most bar refers to teacher performance on truly passing learners where the model predicts with a high confidence that they pass (KK Pass). In this case, 20% of the teachers correctly made the predictions that learners are very likely to pass (solid blue bar), 38% said again correctly that learners are likely to pass (dotted blue bar). 40% wrongly predicted the learners are likely to fail (dotted red bar), and 2% of them incorrectly predicted that learners are very likely to fail (solid red bar). For the sake of comparisons with the machine-learning model, in the x-axis labels, we reported the machine-learning model performance on students belonging to that prediction group (which is, as per our definition, 100% on known-known students, and 0% for the known unknowns and the unknown unknowns students). In the plot title, we reported the overall teacher accuracy and model accuracy, measured by averaging the accuracy values across the corresponding three prediction groups in the plot. For example, in Fig. 4a, we have a model accuracy of 0.33, obtained by averaging its 100% of accuracy in KKPass students, and the two accuracy values of 0% for KUPass and UUPass students. The same rationale has been applied to organize results in Fig. 4b.

Considering the left plot (truly passing learners), it can be observed that the teacher decisions diverged consistently from those of the model. Specifically, the left most bar shows that the teachers did not find as easy as the model to make correct decisions on KK `Pass` learners. While the model was 100% accurate and highly confident about that group of learners, the accuracy of teacher decisions was slightly below 60% (left most blue bar). Regarding the other two groups of learners (mid and right bars), the teachers were more accurate than the model (0% of accuracy). Surprisingly, teachers were overall very good at detecting truly passing learners among those the model was wrong and highly confident (right blue bar). Unknown unknown risks for the model would be prevented once indicators are shown to the teacher. We believe that this is a positive outcome, given the huge amount of unknown unknowns that accurate neural networks applied to educational data tend to produce [13].

Observing the decisions made by teachers on the truly failing learners (right plot), it

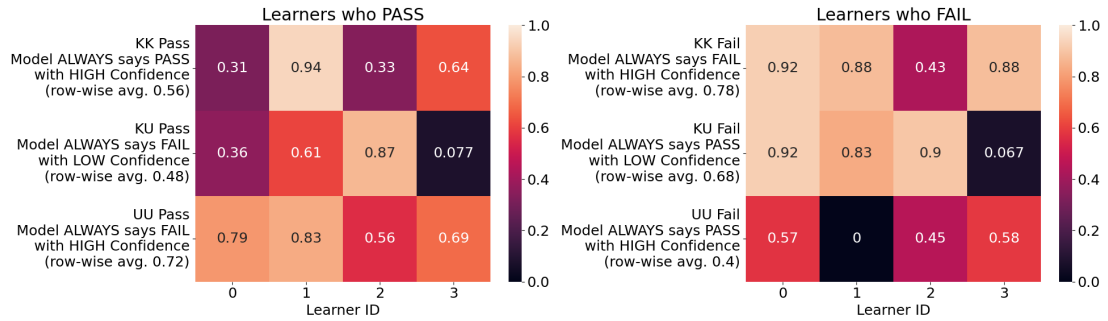


Fig. 5.10: Percentage of teacher decision (RQ1)

conversely emerged that the model and the teacher decisions were more similar. To appreciate this phenomenon, we invite the reader to note the decreasing percentage of correct answers from left to right (blue bars). Both the teachers (79%) and the model (100%) were very accurate on model's known knowns (left bar). On the other hand, teachers predicted well to a good extent (60%) on the learners where the model was wrong and slightly confident (KU `Fail`). Teachers' accuracy (only 40%) further decreased on model's unknown unknowns (model accuracy 0%). Such circumstance is alarming, since both teachers and the model tend to incorrectly predict that a learner would not need an intervention.

Learner-level Correctness We investigated whether the observed patterns reflect a general difficulty of teachers in making correct decisions on learners in that prediction group or emerge solely from individual learners. Fig. 5.10 shows the percentage of correct predictions made by teachers on each learner of each prediction group, referring to truly passing learners (left) and truly failing learners (right). As an example, the top left most cell in the left heatmap indicates that 31% of teachers correctly predicted that the learner with ID 0 in the KK `Pass` group would pass the course. For convenience, we again distinguish patterns between truly passing and failing learners.

From the left heatmap, it emerged that the teacher performance within a prediction group substantially varied across learners. This means that teachers did not generally align with the model. In the first row (KK `Pass`), only learner 1 was characterized by a 94% accuracy, close to that of the model (100%), thus easier to detect for both teachers and the model. On the other hand, teachers found it hard to detect the other truly passing learners (learners 0, 2, and 3). Similar observations can be made on passing known unknowns (KU `Pass`). It should be noted that the model was not able to correctly detect that all those learners would have passed the course. Despite of being underperforming, teachers were overall better than model. Both teachers and model found it challenging to detect that learner 3 would pass. On passing unknown unknowns (UU `Pass`), teachers were remarkably successful about all learners (from 56% to 83%).

Experimental results on truly failing learners (right heatmap) showed again a different behavior. Teachers were overall able to correctly detect learners likely to fail for model's

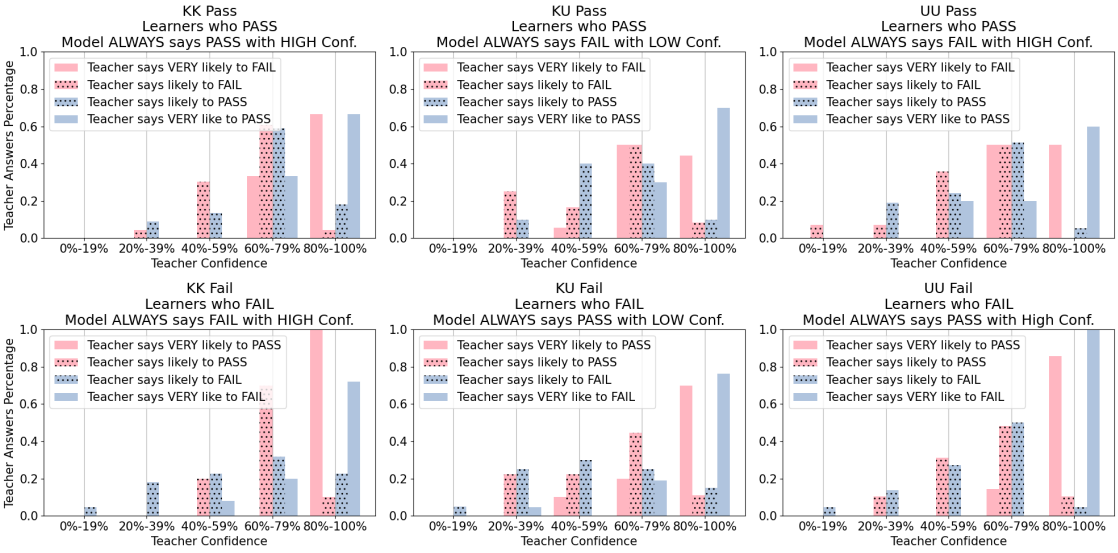


Fig. 5.11: Distribution of teachers' confidence for their prediction (RQ2)

known knowns (KK `Fail`) and known unknowns (KU `Fail`). In each cases, there was a single exception, i.e., learner 2 and learner 3 respectively. Unknown unknowns (UU `Fail`) showed all an accuracy between 40% and 60%, except for learner 1. In the latter group, it was evidently very difficult to identify failures for teachers. Both the teachers and the model incorrectly predicted a passing.

Teachers' Reasoning The barely low performance of teachers on certain learners motivated us to analyze their reasoning. For conciseness, we provide observations on key learners associated to the low accuracy for teachers[4].

For instance, considering the KK `Pass` group, teachers who made the wrong prediction on learner 0 were generally influenced by the decreasing trend in all the indicators across weeks. For instance, teachers said that it, "*Seems like he had the course already passed by the end so he relaxed*", "*I believe that a strong finish is more important than a strong start*", "*They seem to have lost motivation towards the end of the course when they should be working towards a final push.*". Teachers who correctly predict a pass for learner 0 were instead driven by the general estimates, saying, "*This student seems to always be above average at most times*". Another key learner is represented by learner 3 in the KU `Pass` group - only 7.7% of teachers correctly identified them as a passing learner. Teachers who found it challenging to make the right prediction gave equal importance to all indicators and to their up and downs. A teacher reported that, "*The progress bar seems generally similar to students who failed for all/most aspects, except proactivity. Yet the pick up on proactivity might suggest the student has checked what's coming and that has further decreased their engagement with the course.*". Another teacher said that, "*He has had ups and downs related to effort and it seems he/she doesn't have a routine of study.*

[4]The corresponding visualizations for all learners are reported at shorturl.at/FHINT.

And as it is for assessment, he/she hasn't practiced at all during the course. Either he/she is very lucky on the exam or the lack of work will make him/her fail". A teacher who made the right prediction for learner 3 instead reported that, "*The student is engaged and ahead of schedule, absorbs the material without using the pause button and can't be bothered doing the non-compulsory quizzes. I guess that the student has a lot of self-confidence and finds the course easy*".

There were key learners associated to low teacher accuracy also in prediction groups including true failing learners. In the group KU `Fail`, teachers were under-performing on learner 3 particularly. Incorrect predictions were often motivated by the learners' indicator pertaining to assessment, which was showing very low values ("*The results of the students in the assessment is the main factor that made me decide to estimate that the student is likely to pass. The difference between the average passing and the average failing student on the other factors is too small for me to use them to make a decision about the student"* and "*His strong finish in assessment in the final week also means that he seems to be in good shape and that there is no decrease in motivation over time*"). Similar comments were made also on learner 1 within the UU `Fail` group. All teachers made the wrong prediction, reporting, "*He wasn't very good with control, but most of the weeks he was above average on the remaining 3 categories*"", and "*Seem to be working hard towards the end*".

Findings RQ1. *Teachers were overall more accurate in detecting struggling learners. Teachers and model decisions diverged on learners who passed the course. About those who actually failed, teachers were more accurate than the model, but both found the same learners hard to predict. Reasoning aspects that differentiated teachers' correct and incorrect predictions often referred to the importance given to the indicators and the way decreasing/increasing trends were judged.*

RQ2: Correctness-Confidence Relationship

We were then interested in investigating teachers' confidence in their predictions at prediction group- and learner-level.

Group-level Relationship Fig. 5.11 collects the percentage of answers given by teachers under each confidence level. We grouped them according to the answers' correctness. The three bar plots in the top (bottom) row refer to the prediction groups including truly passing (failing) learners. Blue (red) bars refer to teachers who answered correctly (incorrectly). Solid (dotted) bars refer to teacher decisions with "very likely" ("likely"). Bars of the same type sum to 1. As an example, in the top left most bar plot (KK `Pass` group), 62% of the teachers who said the learner would be very likely to pass had a confidence in $[80\%, 100\%)$ (solid blue bar on the right).

Concerning truly passing learners where the model is correct and confident (KK `Pass`), it can be observed that teachers' confidence estimates were high as well. Unfor-

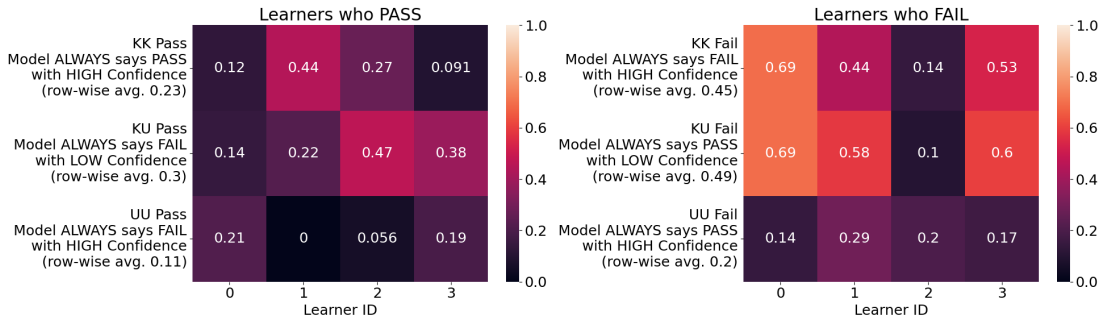


Fig. 5.12: Percentage of teachers showing a confidence (RQ2)

tunately, such high estimates were observed for both correct (blue bars) and incorrect (red bars) predictions, meaning that being confident in their predictions did not imply correctness. We observed a relationship between confidence and correctness for learners in KU `Pass` (top mid bar plot). It emerged that correct predictions were associated to higher confidence (blue bars higher on the right). Incorrect predictions were instead made with a lower confidence. It follows that learners in this group are known unknowns for both the model and the subset of teachers who made the wrong prediction. Similar patterns were reported for learners in UU `Pass`. Except for a small portion of teachers (solid red bar under the 80% - 100% confidence level), incorrect predictions were accompanied by a lower confidence. Being such learners unknown unknowns for the model, we believe this to be a positive outcome of our analysis. Teachers' analysis can help to reduce the cases of learners on which teachers would unnecessarily intervene on, which is important under low teaching resources.

Moving to truly failing learners (bottom row), the left most plot collects confidence distributions on learners the model was both right and confident (KK `Fail`). It can be observed that teachers were very confident regardless of the correctness of their predictions. Compared to truly passing learners, the confidence of teachers in case of a wrong prediction was higher. This observation highlights a high variance across teachers, despite of being presented with the same information for that learner. Overall, although the model found easy to detect failure, this does not apply to teachers - often they were confident but wrong. For KU `Fail` and UU `Fail`, there were no remarkable differences in confidence between wrong and right predictions. This observation is especially alarming for the latter case, where both the model and the teachers who made the wrong prediction were very confident. Consequently, neither the model nor a significant portion of teachers would have confidently identified the need for intervention for those learners.

Learner-level Relationship We were again interested in understanding the extent to which the confidence estimates depend on the prediction group in general or on peculiar individual learners. Fig. 5.12 collects the percentage of teachers who reported a confidence higher than 80% for each learner within each prediction group. Lower values (so

Table 5.5: Teachers' Knowledge and Interventions.

Topic	Examples	KK Pass	KU Pass	UU Pass	KK Fail	KU Fail	UU Fail	Sum	Percentage
Performances in class	In-class participation and performance, attempts in in-class exercises etc.	7	7	11	**14**	9	8	56	14.78
Students background	Demographics, past grades, academic path etc.	9	**10**	9	**10**	8	8	54	14.25
Personal life circumstances	Check with the learner what happened in a specific week	6	6	7	3	4	**8**	34	8.97
Communication with the learner	Reaching out to the learner via a meeting, email, chat etc.	14	16	17	**25**	21	13	106	27.97
Additional support provisioning	Extended tutoring, personalized engagement activities and feedback etc.	**12**	9	**12**	**12**	**12**	**12**	69	18.20
Revision recommendation	Inviting learners to revise a specific lesson or topic	**2**	1	1	1	1	**2**	8	2.11
Learner motivation support	Finding ways to motivate learners to perform better	1	**8**	5	1	5	7	27	7.12

darker cells) indicate learners where teachers were particularly unsure. As an example, in the left heatmap, the left most cell in the first row (KK `Pass` group, learner 0) shows that only 12% of teachers had a confidence higher than 80%.

Considering the truly passing learners (left heatmap), the average percentage of very confident teachers was barely low (30% at most for KU `Pass`). It is interesting to see that the model and the teachers showed an opposite pattern in confidence. Specifically, the learners were the model showed a lower confidence (mid row) were those on which the teachers reported the highest confidence, and viceversa. Within each prediction group, there was a non negligible variance learners. It follows that, although a general trend within each prediction group emerged, there might be peculiar learners who left teachers particularly unsure. Learners in the UU `Pass` group (bottom row) were those where the teachers reported overall the highest accuracy but were not so confident about them.

Teacher predictions on truly failing learners (right heatmap) again led to different observations. Confident predictions were consistently made on the groups KK `Fail` and KU `Fail`. It follows that the teachers and the model were similar on the former group and opposite on the latter group, in terms of confidence. Differently from the model, although being both often wrong, teachers would be aware of the fact that predictions on KU `Fail` learners would require extra care. Such aspect is beneficial to avoid cases of learners missing an intervention actually needed.

Findings RQ2. *Teachers and the model were characterized by a different behavior concerning their confidence in predictions. Teachers tended to be more confident in predicting on truly failing learners than truly passing learners. Overall, on learners where the model was both wrong and confident, teachers showed a low confidence.*

RQ3: Knowledge and Intervention Needs

In a later stage, we conducted an affinity diagramming analysis to complement our findings on correctness and confidence. The aim was to identify any significant needs highlighted by teachers in order to make more informed decisions and determine the actions they would take to address these needs. We delved deeper into the responses provided by teachers to the last two questions in the questionnaire for each learner.

Table 5.5 provides a quantitative summary, and the following are extended examples of their comments.

Considering the *KK Pass* group, some teachers showed particular interest on the assessment estimates - "*it would be relevant to know the specific grades and attempts that student made in the assessment graph*" and the criteria to compute the final grade "*I would like to know how their final grade would be calculated*". Another teacher was interested in knowing more about the learner's life. In particular, they reported that "*It would be good to know a little about the learners private circumstances to understand what might be influencing the sudden drop in engagement*". A teacher also unexpectedly asked for demographic attributes like the gender ("*I would have liked to know the gender of the student*"). As possible interventions, a large group of teachers agreed that it is necessary to talk to the learners, for instance to "*ask to the student if he/she is confident, as he/she stopped practicing with the quizzes*". Others agreed that it is necessary to an take an individual meeting, i.e. "*saying them they need to watch the videos intently and really be clear and precise with the material so pausing and re-watching is very helpful for this and then their assessments should go up*".

In the *KU Pass* group, teachers were in agreement on the need of "*The background of the student grades on related courses (the fact that the student spend less time online, can be connected with they previous knowledge on the topic and on they ability to learn easily the content of the course)*". Some teachers asked themselves several questions, such as "*Is the student pausing a lot because he is taking notes or because he is struggling?*" and "*How the student has performed in previous courses?*". Some teachers made a similar consideration regarding the fact that they should encourage the learner to engage more with the online content, e.g., "*ask the student to not skip any voluntary quiz*" and "*e-mail them to check*".

Regarding the *UU Pass* group, teachers agreed with the importance of having grades, such as "*It's important to know the results of tests that count to the grade during the course*". Another important common view is the importance of knowing the learners personally, i.e. "*If I knew the student more personally I would perhaps have a better guess or at least feel more confident in my guess*" and knowing extra information, "*Maybe if I knew how the student did in other courses with the same data that is being presented here I would know if the student usually passes with ease or is just having some trouble with this course*". Teachers thought that it is also important to make comparisons of data taken a year before to see the increase or decline in the students performance. An important way to intervene is to "*discuss with the student to watch the videos properly with no distractions and have them check in explaining what was happening*".

In the *KK Fail* group, teachers agreed that "Knowing the student performance in class, such as results in graded exams and previous years grades would help". A common idea is that the more you know about learners' background, the more you will be able to help them, i.e. "*To know if any family or personal considerations have influenced the performance*" and "*If the student had any learning difficulties or trouble working with technology such as computers*". Another teacher suggest that "*It would be good to know if they at-*

tended the class and attempting the quizzes where the student scored zero. This would rule out the student already being competent in this course". Knowing more about performance is a crucial key, i.e. "*I think it would be good to know how they are performing during 'in-class' as they may have natural ability which is not picked up by the analysis of the 'pre-class' tasks*". Talking with the student is another common idea.

In the *KU Fail* group, different teachers were interested in knowing more about personal student life, i.e. "*know psycho-social aspects about their lack of interest*", or "*How is the student's home life, who do they live with?*". An important observation was related to the latter part of the course, which can clearly be challenging, "*In future I would perhaps forewarn all the students about it before the beginning of the course, and open some discussion channels they can post questions outside of class. I would also arrange for more practice material*". Sometimes it can be necessary to offer individual plan for the student, i.e. "*if they are not interested in the course, maybe a different teaching/learning approach might help, if they are advanced, further development might provide stimulation and keep them interested*".

Finally, in the *UU Fail* group, some teachers agreed that "*I should have known the academic path of the student*". Others thought that it is necessary to motivate the learners, e.g., "*try to motivate they by discussing what might be interesting aspects of the course, and offer some form of tutoring*" and to support them, "*I would offer them some tutor sessions so they can get back on track*", "*Suggest the student to be more constant*", and "*I would encourage them to demonstrate their understanding during class and offer interventions if it was weak*".

Findings RQ3. *Despite the variety of the considered learners and of the teachers population, teachers overall highlighted similar aspects. The level of participation and of the performances in class emerged as very important for teachers. In addition, knowledge about the learners' life and prior courses performance were deemed as important.*

5.2.4 Findings and Recommendations

The results presented in Section 5.2.3 (**RQ1**) demonstrate that teachers are overall more accurate in detecting struggling learners. Tasks in education still represent a challenge for machines. Education therefore differ from other fields like computer vision, where models surpassed humans in performance under several tasks, e.g., [169]. Furthermore, teachers and model decisions diverged on learners who passed the course. It means that they found difficult to correctly estimate that learners would pass the course on highly disjoint populations. We therefore found that our observations agree with those of other studies in the neuroscience field attempting to investigate how humans and models differ in perception [170]. Future work should investigate novel methods to let both teachers and the model giving/receiving feedback from each other and adapting their predictions accordingly. This line of research has been just recently experimented

in education, with the creation of explainable models. On learners who actually failed, both teachers and model found, on average, the same learners hard to predict. It follows that both of them had the tendency to over-estimate learners' capabilities of passing the course and therefore potentially miss actually needed interventions. For some teachers, additional contextual information would be needed to better inform their decisions, in agreement with studies like [166]. Future work should be devoted to devising tracking methods for in-class activities and modelling behavioral indicators on top. Prior work has found that models close to human understanding also generalise better [171]. We believe that the overall low performance of both this tasks might be also present in other courses, therefore requiring further investigation. However, future work should extend this study to several courses of different types, to better investigate the reproducibility power of the results in larger contexts. The second research question (**RQ2**) was designed to investigate relationships between correctness and confidence in predictions. Decision confidence generally reflects teachers' ability to evaluate the quality of decisions and guides subsequent behavior. The results presented in Section 5.2.3 suggest that teachers and the model were characterized by a different confidence behavior. Teachers tended to have a limited confidence in their decisions regardless of the prediction group, except for cases concerning truly failing learners. Performance confidence in humans is complex and influenced also by physical properties of the stimulus that a decision is based on. For example, the quality of evidence favoring a decision has been shown to affect confidence [172]. We therefore conjecture that the observed low confidence estimates might be influenced by the generally low confidence of teachers in using technologies [173]. Building on our findings, future work should go deeper in understanding the interdependencies between teachers' confidence and factors such as the displayed graphs (e.g., their type), the selected indicators, the teacher capability in interpreting graphs and so on. Overall, confidence and correctness in decisions tended to be more aligned in teachers than the model. Future research from the machine learning perspective should therefore devise novel methods to prevent the risk of unknown unknowns [13].

Finally, the results presented in Section 5.2.3 provide evidence to answer the third research question (**RQ3**). They highlighted that, despite the variety of learners considered, teachers overall pointed out to similar aspects. One key element emerged from our study is that learning indicators in the current course should be accompanied by a short summary of the learner' profile. It emerged that such profile description should include, as examples, performance in prior courses, demographic information, and any current personal issue. However, collecting and reporting information on the above subjects opens to key privacy issues which might not often justify their final scope [174]. It becomes therefore urgent to devise privacy-aware methods that allow to better contextualize the learning situation and clarify the cause-effect links that led to passing or failing a course, without affecting learners' privacy. Indeed, being not aware of certain information, it is one of the key factors that lead both the teacher and the model to face unknown risks. Future work should also therefore investigate which currently hidden

variables caused the reported wrong predictions.

Our study investigates how close machine-learned models are to teachers at predicting learners at risk. Student success prediction models are an essential part of the methodology that enables the creation of a student model, being performance a major factor that allows a platform to personalize the student learning experience. Our results in this paper can be useful to be considered when building student models with this aim. In particular, the results highlight some potential (dis)agreements in the predictions between machine-learning models and teachers and uncover issues that can arise in case such models are used to build better student models. Our findings therefore make it evident the urgent need of re-considering the use of student success prediction models for student model creation, since the underlying performance especially on unknown unknown students might lead to undesired and wrong personalization strategies according to the resulting student model. Future work will also embrace the current limitations to show generalization of our findings on a larger set of models and courses. Our findings showcase the potential of making teachers and models collaborate for detecting learners at risk, serving as an important point for effective augmented intelligence in personalized education. In order to design future works, it should be considered that teachers may lack information about students, since these are not their students, and the visualizations they need to analyze might lack important data (e.g. in some cases lines for average passing and and average failing students almost coincide). However, it should be considered that the students shown to teachers are actually those where the model might lead to undesired outcomes (they are hard cases). Even though teachers do not always decide to intervene only based on plots, there are several contexts where they have a limited view of the students' learning process and, therefore, such diagrams and predictions are the main source to rely on.

Chapter 6

Conclusions

In this thesis, we addressed the critical issue of responsibility in AI-enabled educational systems and its impact on student learning experiences. We began by investigating the challenges and needs faced by the educational community in integrating responsibility into AI applications. Moreover, we explored algorithmic disparities in both synchronous and asynchronous learning environments. Additionally, we delved into student success prediction and the presence of unknown unknowns in educational models, shedding light on the limitations and complexities of student success prediction.

6.1 Contribution Summary

Throughout this doctoral thesis, we have made significant contributions to the field of AI in education, with a specific focus on responsibility, inclusivity, and improved learning experiences for all students. Our research aimed to address the challenges and needs in developing responsible AI-enabled educational systems by investigating various aspects of algorithmic disparities and student behavior analysis.

Firstly, we conducted an in-depth exploration of expert views on responsibility in AI for education. Through a survey and semi-structured interviews with educational researchers and practitioners, we gained valuable insights into their perspectives, challenges, and priorities regarding responsible-aware practices. This comprehensive understanding of experts' opinions served as a solid foundation for our subsequent research.

Secondly, we delved into the analysis of algorithmic disparities in synchronous and asynchronous learning environments. We employed clustering techniques to identify participation patterns in synchronous learning, shedding light on how student behavior impacts learning outcomes. Additionally, we investigated the presence of "unknown unknowns" in student success prediction models, uncovering the complexities and limitations of these models in predicting student outcomes accurately.

6.2 Take-home Messages

Throughout this research journey, several key take-home messages emerged, guiding the future development of responsible AI-powered educational systems. Firstly, the integration of responsible-aware practices should be prioritized when designing and deploying AI applications in the educational landscape. Domain-specific resources, metrics, and processes are essential to address the unique challenges that emerge in various educational contexts. Secondly, the analysis of student behavior in both synchronous and asynchronous learning modes holds profound implications for optimizing teaching strategies, enhancing student engagement, preventing disengagement, and predicting learning outcomes. This analysis represents a crucial stepping stone towards personalizing education and tailoring interventions to meet individual student needs. Thirdly, acknowledging and addressing unknown unknowns in student success prediction models are paramount to ensure the accuracy, reliability, and ethical responsibility of AI interventions in education. Embracing uncertainties and model limitations will lead to more effective and trustworthy predictive models.

6.3 Future Research Directions

This doctoral thesis has opened up several promising avenues for future research in the field of AI in education. As we move forward, the following research directions are worth exploring:

- **Context-specific Fairness**. Investigating fairness considerations in diverse educational contexts, such as different grade levels, subject domains, and cultural backgrounds, to ensure that fairness-aware practices are adaptable and effective in various settings.
- **Human-in-the-loop Decision-making**. Exploring the integration of human decision-making with AI predictions to address uncertainties and unknowns. A collaborative approach that combines human expertise with machine intelligence can lead to more robust and reliable educational AI systems.
- **Enhanced Student Success Prediction Models**. Advancing student success prediction models by incorporating additional factors, such as students' prior knowledge, socio-economic backgrounds, and metacognitive skills, to improve their accuracy and predictive power.
- **Longitudinal Analysis**. Conducting longitudinal studies to gain a deeper understanding of how student behavior evolves over time and how it impacts long-term educational outcomes.
- **Ethical Frameworks for AI in Education**. Developing ethical frameworks and guidelines for the responsible design, deployment, and evaluation of AI-enabled

educational systems, ensuring that the potential risks and benefits are carefully considered and balanced.

- **Integration of ChatGPT and Metaverse in Education**. Recognizing the emerging trends in technology, it is important to consider the potential impact of chat-based AI systems, such as ChatGPT, and immersive technologies like the Metaverse in the realm of education. While the primary focus of this thesis has been on specific aspects of AI in education, future research could explore the integration of ChatGPT for personalized learning experiences, interactive tutoring, and student engagement. Additionally, the utilization of the Metaverse as a platform for collaborative and immersive educational environments deserves attention.

In conclusion, this thesis represents a significant stride towards promoting responsibility, inclusivity, and ethical responsibility in AI-powered educational systems. By addressing the challenges and needs of the educational community, analyzing algorithmic disparities in learning environments, and unraveling the intricacies of student success prediction, we establish a strong foundation for future research endeavors. As we continue to push the boundaries of AI in education, it is crucial to prioritize responsibility, transparency, and the human aspect of learning to create a transformative and equitable educational landscape for all students.

References

[1] J. L. Rastr.-Guerr., J. A. Gomez-Pulido, and A. Duran-Dom., "Analyzing and predicting students' perf. by means of ml: a review," *Appl. sciences*, vol. 10, no. 3, p. 1042, 2020.

[2] A. Abyaa, M. Khalidi Idrissi, and S. Bennani, "Learner mod.: systematic review of the lit. from the last 5 years," *Educ. Tech. Res. and Dev.*, vol. 67, no. 5, pp. 1105–1143, 2019.

[3] S. M. R. Abidi, M. Hussain, Y. Xu, and W. Zhang, "Prediction of confusion attempting algebra homework in an intelligent tutoring system through machine learning techniques for educational sustainable development," *Sustainability*, vol. 11, no. 1, p. 105, 2019.

[4] T. Alsuliman, D. Humaidan, and L. Sliman, "Ml and ai in the service of medicine: Necessity or potentiality?," *Current res. in translational med.*, vol. 68, no. 4, pp. 245–251, 2020.

[5] J. Wang, M. D. Molina, and S. S. Sundar, "When expert recommendation contradicts peer opinion: Relative social influence of valence, group identity and artificial intelligence," *Computers in Human Behavior*, vol. 107, p. 106278, 2020.

[6] J. Britto, S. Prabhu, A. Gawali, and Y. Jadhav, "A machine learning based approach for recommending courses at graduate level," in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 117–121, IEEE, 2019.

[7] Z. Shahbazi and Y. C. Byun, "Toward social media content recommendation integrated with data science and ml approach for e-learners," *Symmetry*, vol. 12, no. 11, p. 1798, 2020.

[8] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*, pp. 1–7, IEEE, 2018.

[9] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proceedings of the 22nd ACM SIGKDD int. conference on knowledge discovery and data mining*, pp. 2125–2126, 2016.

[10] G. Fenu, R. Galici, and M. Marras, "Experts' view on challenges and needs for fairness in artificial intelligence for education," in *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 243–255, Springer, 2022.

[11] G. Fenu and R. Galici, "Modelling student behavior in synchronous online learning during the covid-19 pandemic.," in *L2D@ WSDM*, pp. 28–40, 2021.

[12] G. Fenu, R. Galici, M. Marras, and S. Picciau, "Supporting instructors with course attendance and quality prediction in synchronous learning," in *International Workshop on Higher Education Learning Methodologies and Technologies Online*, pp. 71–83, Springer, 2022.

[13] R. Galici, T. Kaser, G. Fenu, and M. Marras, "Do not trust a model because it is confident: Uncovering and characterizing unknown unknowns to student success predictors in online-based learning," in *LAK23: 13th International Learning Analytics and Knowledge Conference*, pp. 441–452, 2023.

[14] R. Galici, T. Käser, G. Fenu, and M. Marras, "How close are predictive models to teachers in detecting learners at risk?," in *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pp. 135–145, 2023.

[15] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[16] B. Apolloni, A. Ghosh, F. Alpaslan, and S. Patnaik, *Machine learning and robot perception*, vol. 7. Springer Science & Business Media, 2005.

[17] S.-I. Ao, B. B. Rieger, and M. Amouzegar, *Machine learning and systems engineering*, vol. 68. Springer Science & Business Media, 2010.

[18] L. Györfi, G. Ottucsák, and H. Walk, "Machine learning for financial engineering, 2012."

[19] J. Yu and D. Tao, *Modern machine learning techniques and their applications in cartoon animation research*. John Wiley & Sons, 2013.

[20] Y. Gong and W. Xu, *Machine learning for multimedia content analysis*, vol. 30. Springer Science & Business Media, 2007.

[21] A. Fielding, *Machine learning methods for ecological applications*. Springer Science & Business Media, 1999.

[22] S. Mitra, S. Datta, T. Perkins, and G. Michailidis, *Introduction to machine learning and bioinformatics*. CRC Press, 2008.

[23] Z. R. Yang, *Machine learning approaches to bioinformatics*, vol. 4. World scientific, 2010.

[24] T. J. Cleophas, A. H. Zwinderman, and H. I. Cleophas-Allers, *Machine learning in medicine*, vol. 9. Springer, 2013.

[25] J. D. Malley, K. G. Malley, and S. Pajevic, *Statistical learning for biomedical data*. Cambridge University Press, 2011.

[26] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[27] T. M. Mitchell, "Machine learning," 1997.

[28] S. Maghsudi, A. Lan, J. Xu, and M. van Der Schaar, "Personalized education in the artificial intelligence era: what to expect next," *IEEE Signal Processing Magazine*, vol. 38, no. 3, pp. 37–50, 2021.

[29] D. Shawky and A. Badawi, "Towards a personalized learning experience using reinforcement learning," *Machine learning paradigms: Theory and application*, pp. 169–187, 2019.

[30] M. Ross, C. A. Graves, J. W. Campbell, and J. H. Kim, "Using support vector machines to classify student attentiveness for the development of personalized learning systems," in *2013 12th international conference on machine learning and applications*, vol. 1, pp. 325–328, IEEE, 2013.

[31] T. Doleck, D. J. Lemay, R. B. Basnet, and P. Bazelais, "Predictive analytics in education: a comparison of deep learning frameworks," *Education and Information Technologies*, vol. 25, pp. 1951–1963, 2020.

[32] S. D. A. Bujang, A. Selamat, and O. Krejcar, "A predictive analytics model for students grade prediction by supervised machine learning," in *IOP Conference Series: Materials Science and Engineering*, vol. 1051, p. 012005, IOP Publishing, 2021.

[33] V. L. Uskov, J. P. Bakken, A. Byerly, and A. Shah, "Machine learning-based predictive analytics of student academic performance in stem education," in *2019 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1370–1376, IEEE, 2019.

[34] A. Alam, "A digital game based learning approach for effective curriculum transaction for teaching-learning of artificial intelligence and machine learning," in *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pp. 69–74, IEEE, 2022.

[35] M. Tedre, T. Toivonen, J. Kahila, H. Vartiainen, T. Valtonen, I. Jormanainen, and A. Pears, "Teaching machine learning in k–12 classroom: Pedagogical and technological trajectories for artificial intelligence education," *IEEE access*, vol. 9, pp. 110558–110572, 2021.

[36] K. Fahd, S. Venkatraman, S. J. Miah, and K. Ahmed, "Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature," *Education and Information Technologies*, pp. 1–33, 2022.

[37] J. Luan, "Data mining applications in higher education," *SPSS Executive*, vol. 7, 2004.

[38] B. P. Woolf, *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann, 2010.

[39] M. Ciolacu, A. F. Tehrani, R. Beer, and H. Popp, "Education 4.0—fostering student's performance with machine learning methods," in *2017 IEEE 23rd international symposium for design and technology in electronic packaging (SIITME)*, pp. 438–443, IEEE, 2017.

[40] F. Ofori, E. Maina, and R. Gitonga, "Using machine learning algorithms to predict students' performance and improve learning outcome: A literature based review," *Journal of Information and Technology*, vol. 4, no. 1, pp. 33–55, 2020.

[41] W. Xing and D. Du, "Dropout prediction in moocs: Using deep learning for personalized intervention," *Journal of Educational Computing Research*, vol. 57, no. 3, pp. 547–570, 2019.

[42] B. Mahakud, B. Parida, I. Panda, S. Maity, A. Sahoo, and R. Sharma, "A machine learning system to monitor student progress in educational institutes," *arXiv preprint arXiv:2211.05829*, 2022.

[43] S. Hussain and M. Q. Khan, "Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning," *Annals of data science*, vol. 10, no. 3, pp. 637–655, 2023.

[44] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan, "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," *Ieee Access*, vol. 9, pp. 7519–7539, 2021.

[45] R. Al-Shabandar, A. J. Hussain, P. Liatsis, and R. Keight, "Detecting at-risk students with early interventions using machine learning techniques," *IEEE Access*, vol. 7, pp. 149464–149478, 2019.

[46] M. Tan and P. Shao, "Prediction of student dropout in e-learning program through the use of machine learning method.," *International journal of emerging technologies in learning*, vol. 10, no. 1, 2015.

[47] D. Tiene, "Online discussions: A survey of advantages and disadvantages compared to face-to-face discussions," *Journal of educational multimedia and hypermedia*, vol. 9, no. 4, pp. 369–382, 2000.

[48] S. Bali and M. Liu, "Students' perceptions toward online learning and face-to-face learning courses," in *Journal of Physics: conference series*, vol. 1108, p. 012094, IOP Publishing, 2018.

[49] V. Gherheș, C. E. Stoian, M. A. Fărcașiu, and M. Stanici, "E-learning vs. face-to-face learning: Analyzing students' preferences and behaviors," *Sustainability*, vol. 13, no. 8, p. 4381, 2021.

[50] A. R. Artino Jr, "Online or face-to-face learning? exploring the personal factors that predict students' choice of instructional format," *The Internet and Higher Education*, vol. 13, no. 4, pp. 272–276, 2010.

[51] Y. J. Park and C. J. Bonk, "Synchronous learning experiences: Distance and residential learners' perspectives in a blended graduate course," *Journal of Interactive Online Learning*, vol. 6, no. 3, pp. 245–264, 2007.

[52] Q. Wang, C. L. Quek, and X. Hu, "Designing and improving a blended synchronous learning environment: An educational design research," *The International Review of Research in Open and Distributed Learning*, vol. 18, no. 3, 2017.

[53] O. Marjanovic, "Learning and teaching in a synchronous collaborative environment," *Journal of Computer Assisted Learning*, vol. 15, no. 2, pp. 129–138, 1999.

[54] E. Szeto and A. Y. Cheng, "Towards a framework of interactions in a blended synchronous learning environment: what effects are there on students' social presence experience?," *Interactive Learning Environments*, vol. 24, no. 3, pp. 487–503, 2016.

[55] Y. Kim, B. Choi, S. Kang, B. Kim, and H. Yun, "Comparing the effects of direct and indirect synchronous written corrective feedback: Learning outcomes and students' perceptions," *Foreign Language Annals*, vol. 53, no. 1, pp. 176–199, 2020.

[56] N.-S. Chen, H.-C. Ko, Kinshuk*, and T. Lin, "A model for synchronous learning using the internet," *Innovations in Education and Teaching International*, vol. 42, no. 2, pp. 181–194, 2005.

[57] M. Bower, B. Dalgarno, G. E. Kennedy, M. J. Lee, and J. Kenney, "Design and implementation factors in blended synchronous learning environments: Outcomes from a cross-case analysis," *Computers & Education*, vol. 86, pp. 1–17, 2015.

[58] L. C. Yamagata-Lynch, "Blending online asynchronous and synchronous learning," *International Review of Research in Open and Distributed Learning*, vol. 15, no. 2, pp. 189–212, 2014.

[59] R. Wegerif, "The social dimension of asynchronous learning networks," *Journal of asynchronous learning networks*, vol. 2, no. 1, pp. 34–49, 1998.

[60] C. B. Hodges, "Self-regulation of learners in an asynchronous university math course," *Quarterly Review of Distance Education*, vol. 10, no. 2, pp. 233–237, 2009.

[61] E. Delen and J. Liew, "The use of interactive environments to promote self-regulation in online learning: A literature review.," *European Journal of Contemporary Education*, vol. 15, no. 1, pp. 24–33, 2016.

[62] V. Sher, M. Hatala, and D. Gasevic, "Analyzing the consistency in within-activity learning patterns in blended learning," ACM, 2020.

[63] P. Mejia-Domenzain, M. Marras, C. Giang, and T. Käser, "Identifying and comparing multidimensional student profiles across flipped classrooms," vol. 13355, pp. 90–102, Springer, 2022.

[64] G. Northey, T. Bucic, M. Chylinski, and R. Govind, "Increasing student engagement using asynchronous learning," *Journal of Marketing Education*, vol. 37, no. 3, pp. 171–180, 2015.

[65] T. Henderson, "Classroom assessment techniques in asynchronous learning networks," *The Technology Source*, 2001.

[66] J. Gardner and C. Brooks, "Student success prediction in moocs," *User Modeling and User-Adapted Interaction*, vol. 28, no. 2, pp. 127–203, 2018.

[67] Z. Kovacic, "Early prediction of student success: Mining students' enrolment data.," 2010.

[68] A. Shemshack and J. M. Spector, "A systematic literature review of personalized learning terms," *Smart Learning Environments*, vol. 7, no. 1, pp. 1–20, 2020.

[69] J. F. Pane, E. D. Steiner, M. D. Baird, and L. S. Hamilton, "Continued progress: Promising evidence on personalized learning.," *Rand Corporation*, 2015.

[70] J. E. Beck and B. P. Woolf, "High-level student modeling with machine learning," in *International Conference on Intelligent Tutoring Systems*, pp. 584–593, Springer, 2000.

[71] C. Yang, F.-K. Chiang, Q. Cheng, and J. Ji, "Machine learning-based student modeling methodology for intelligent tutoring systems," *Journal of Educational Computing Research*, vol. 59, no. 6, pp. 1015–1035, 2021.

[72] X. Wang and Y. Wang, "Improving content-based and hybrid music recommendation using deep learning," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 627–636, 2014.

[73] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," *Advances in neural information processing systems*, vol. 28, 2015.

[74] I. T. Sanusi, S. S. Oyelere, H. Vartiainen, J. Suhonen, and M. Tukiainen, "A systematic review of teaching and learning machine learning in k-12 education," *Education and Information Technologies*, vol. 28, no. 5, pp. 5967–5997, 2023.

[75] D. Spikol, E. Ruffaldi, G. Dabisias, and M. Cukurova, "Supervised machine learning in multimodal learning analytics for estimating success in project-based learning," *Journal of Computer Assisted Learning*, vol. 34, no. 4, pp. 366–377, 2018.

[76] G. Zhang, "A study of grammar analysis in english teaching with deep learning algorithm," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 15, no. 18, pp. 20–30, 2020.

[77] S. Cunningham-Nelson, W. Boles, L. Trouton, and E. Margerison, "A review of chatbots in education: practical steps forward," in *30th annual conference for the australasian association for engineering education (AAEE 2019): educators becoming agents of change: innovate, integrate, motivate*, pp. 299–306, Engineers Australia, 2019.

[78] B. Berendt, A. Littlejohn, and M. Blakemore, "Ai in education: learner choice and fundamental rights," *Learning, Media and Technology*, vol. 45, no. 3, pp. 312–324, 2020.

[79] A. Renz and R. Hilbig, "Prerequisites for artificial intelligence in further education: identification of drivers, barriers, and business models of educational technology companies," *Int. Journal of Educ. Tech. in Higher Education*, vol. 17, no. 1, pp. 1–21, 2020.

[80] S. Caton and C. Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.

[81] V. Iosifidis, B. Fetahu, and E. Ntoutsi, "Fae: A fairness-aware ensemble framework," in *2019 IEEE Int. Conference on Big Data (Big Data)*, pp. 1375–1380, IEEE, 2019.

[82] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions," *arXiv preprint arXiv:1811.07867*, 2018.

[83] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in ml," *ACM Comp. Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[84] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?," in *Proc. of the 2019 CHI conf. on human factors in comp. systems*, pp. 1–16, 2019.

[85] M. Veale, M. Van Kleek, and R. Binns, "Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making," in *Proceedings of the 2018 chi conference on human factors in computing systems*, pp. 1–14, 2018.

[86] J. Liu and J. Eagan, "Adqda: A cross-device affinity diagramming tool for fluid and holistic qualitative data analysis," *Proc. of the ACM on HC Int.*, vol. 5, no. ISS, pp. 1–19, 2021.

[87] S. Hrastinski, "Asynchronous and synchronous e-learning," *Educause quarterly*, vol. 31, no. 4, pp. 51–55, 2008.

[88] A. T. Peterson, P. N. Beymer, and R. T. Putnam, "Synchronous and asynchronous discussions: Effects on cooperation, belonging, and affect.," *Online Learning*, vol. 22, no. 4, pp. 7–25, 2018.

[89] G. Fenu and M. Marras, "Leveraging continuous multi-modal authentication for access control in mobile cloud environments," in *New Trends in Image Analysis and Processing - ICIAP 2017 - ICIAP International Workshops*, vol. 10590 of *Lecture Notes in Computer Science*, pp. 331–342, Springer, 2017.

[90] S. Barra, M. Marras, and G. Fenu, "Continuous authentication on smartphone by means of periocular and virtual keystroke," in *Network and System Security - 12th International Conference, NSS 2018*, vol. 11058 of *Lecture Notes in Computer Science*, pp. 212–220, Springer, 2018.

[91] F. Yang and F. W. Li, "Study on student performance estimation, student progress analysis, and student potential prediction based on data mining," *Computers & Education*, vol. 123, pp. 97–108, 2018.

[92] M. Cantabella, R. Martínez-España, B. Ayuso, J. A. Yáñez, and A. Muñoz, "Analysis of student behavior in learning management systems through a big data framework," *Future Generation Computer Systems*, vol. 90, pp. 262–272, 2019.

[93] T. Lubicz-Nawrocka and K. Bunting, "Student perceptions of teaching excellence: an analysis of student-led teaching award nomination data," *Teaching in Higher Education*, vol. 24, no. 1, pp. 63–80, 2019.

[94] Y. Wang, W. M. White, and E. Andersen, "Pathviewer: Visualizing pathways through student data," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 960–964, 2017.

[95] P. Prinsloo, S. Slade, and M. Khalil, "Student data privacy in moocs: A sentiment analysis," *Distance Education*, vol. 40, no. 3, pp. 395–413, 2019.

[96] L. R. Amir, I. Tanti, D. A. Maharani, Y. S. Wimardhani, V. Julia, B. Sulijaya, and R. Puspitawati, "Student perspective of classroom and distance learning during covid-19 pandemic in the undergraduate dental study program universitas indonesia," *BMC medical education*, vol. 20, no. 1, pp. 1–8, 2020.

[97] C. G. Brinton, S. Buccapatnam, M. Chiang, and H. V. Poor, "Mining mooc clickstreams: Video-watching behavior vs. in-video quiz performance," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3677–3692, 2016.

[98] G. Kőrösi and R. Farkas, "Mooc performance prediction by deep learning from raw clickstream data," in *International Conference on Advances in Computing and Data Sciences*, pp. 474–485, Springer, 2020.

[99] J. L. Harvey and S. A. Kumar, "A practical model for educators to predict student performance in k-12 education using machine learning," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 3004–3011, IEEE, 2019.

[100] R. I. Rashu, N. Haq, and R. M. Rahman, "Data mining approaches to predict final grade by overcoming class imbalance problem," in *2014 17th International Conference on Computer and Information Technology (ICCIT)*, pp. 14–19, IEEE, 2014.

[101] Q. Cao, T. E. Griffin, and X. Bai, "The importance of synchronous interaction for student satisfaction with course web sites," *Journal of Information Systems Education*, vol. 20, no. 3, p. 331, 2009.

[102] D. Leeder, H. Wharrad, and T. BChir, "Beyond institutional boundaries: reusable learning objects for multi- professional education," 01 2002.

[103] A. Francescucci and L. Rohani, "Exclusively synchronous online (viri) learning: The impact on student performance and engagement outcomes," *Journal of marketing Education*, vol. 41, no. 1, pp. 60–69, 2019.

[104] A. Francescucci and M. Foster, "The viri (virtual, interactive, real-time, instructor-led) classroom: The impact of blended synchronous online courses on student performance, engagement, and satisfaction.," *Canadian Journal of Higher Education*, vol. 43, no. 3, pp. 78–91, 2013.

[105] Ü. Çakıroğlu and S. Kılıç, "Understanding community in synchronous online learning: do perceptions match behaviours?," *Open Learning: The Journal of Open, Distance and e-Learning*, vol. 35, no. 2, pp. 105–121, 2020.

[106] X. Yang, D. Li, X. Liu, and J. Tan, "Learner behaviors in synchronous online prosthodontic education during the 2020 covid-19 pandemic," *The Journal of prosthetic dentistry*, 2020.

[107] T. C. Shoepe, J. F. McManus, S. E. August, N. L. Mattos, T. C. Vollucci, and P. R. Sparks, "Instructor prompts and student engagement in synchronous online nutrition classes," *American Journal of Distance Education*, vol. 34, no. 3, pp. 194–210, 2020.

[108] O. B. Adedoyin and E. Soykan, "Covid-19 pandemic and online learning: the challenges and opportunities," *Interactive Learning Environments*, pp. 1–13, 2020.

[109] A. Bozkurt and R. C. Sharma, "Emergency remote teaching in a time of global crisis due to coronavirus pandemic," *Asian Journal of Distance Education*, vol. 15, no. 1, pp. i–vi, 2020.

[110] J. Feldman, "To grade or not to grade," *Educational Leadership*, vol. 77, no. 10, pp. 43–46, 2020.

[111] O. Scheuer and B. M. McLaren, *Educational Data Mining*, pp. 1075–1079. Boston, MA: Springer US, 2012.

[112] D. Dessì, M. Dragoni, G. Fenu, M. Marras, and D. R. Recupero, "Evaluating neural word embeddings created from online course reviews for sentiment analysis," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC 2019*, pp. 2124–2127, ACM, 2019.

[113] L. Boratto, G. Fenu, and M. Marras, "Connecting user and item perspectives in popularity debiasing for collaborative recommendation," *Inf. Process. Manag.*, vol. 58, no. 1, p. 102387, 2021.

[114] D. Dessì, G. Fenu, M. Marras, and D. R. Recupero, "Leveraging cognitive computing for multi-class classification of e-learning videos," in *The Semantic Web: ESWC 2017 Satellite Events - ESWC 2017 Satellite Events, Portorož*, vol. 10577 of *Lecture Notes in Computer Science*, pp. 21–25, Springer, 2017.

[115] G. Fenu, M. Marras, and M. Meles, "A learning analytics tool for usability assessment in moodle environments," *Journal of e-Learning and Knowledge Society*, vol. 13, no. 3, 2017.

[116] J. Byun, D. Pennington, J. Cardenas, S. Dutta, and J. Kirwan, "Understanding student behaviors in online classroom: data scientific approach," in *2014 IEEE International Congress on Big Data*, pp. 802–803, IEEE, 2014.

[117] J.-L. Hung and K. Zhang, "Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching," *MERLOT Journal of Online Learning and Teaching*, 2008.

[118] J.-L. Hung and S. M. Crooks, "Examining online learning patterns with data mining techniques in peer-moderated and teacher-moderated courses," *Journal of Educational Computing Research*, vol. 40, no. 2, pp. 183–210, 2009.

[119] R. Ahuja, A. Jha, R. Maurya, and R. Srivastava, "Analysis of educational data mining," in *Harmony Search and Nature Inspired Optimization Algorithms*, pp. 897–907, Springer, 2019.

[120] M. W. Rodrigues, S. Isotani, and L. E. Zarate, "Educational data mining: A review of evaluation process in the e-learning," *Telematics and Informatics*, vol. 35, no. 6, pp. 1701–1717, 2018.

[121] S. Kausar, X. Huahu, I. Hussain, Z. Wenhao, and M. Zahid, "Integration of data mining clustering approach in the personalized e-learning system," *IEEE Access*, vol. 6, pp. 72724–72734, 2018.

[122] D. Ding, J. Li, H. Wang, and Z. Liang, "Student behavior clustering method based on campus big data," in *2017 13th International Conference on Computational Intelligence and Security (CIS)*, pp. 500–503, IEEE, 2017.

[123] M. Durairaj and C. Vijitha, "Educational data mining for prediction of student performance using clustering algorithms," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 4, pp. 5987–5991, 2014.

[124] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education*, vol. 113, pp. 177–194, 2017.

[125] M. Beerkens, "Evidence-based policy and higher education quality assurance: progress, pitfalls and promise," *Eur. J. of Higher Education*, vol. 8, no. 3, pp. 272–287, 2018.

[126] A. A. Al-Imarah and R. Shields, "Moocs, disruptive innovation and the future of higher education: A conceptual analysis," *Innovations in Education and Teaching International*, vol. 56, no. 3, pp. 258–269, 2019.

[127] L. Jiang and X. Wang, "Optimization of online teaching quality evaluation model based on hierarchical pso-bp neural network," *Complexity*, vol. 2020, 2020.

[128] J. Chen, H. Li, W. Wang, W. Ding, G. Y. Huang, and Z. Liu, "A multimodal alerting system for online class quality assurance," in *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 381–385, Springer, 2019.

[129] M. Xu, N. Wang, S. Gong, H. Zhang, Z. Zhang, and S. Liu, "Course quality evaluation based on deep neural network," in *Proceedings of the International Conference in Communications, Signal Processing, and Systems*, pp. 56–62, Springer, 2022.

[130] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," *arXiv preprint arXiv:1606.06364*, 2016.

[131] M.-H. Cho and D. Shen, "Self-regulation in online learning," *Distance education*, vol. 34, no. 3, pp. 290–301, 2013.

[132] M. Marras, J. T. T. Vignoud, and T. Käser, "Can feature predictive power generalize? benchmarking early predictors of student success across flipped and online courses," International Educational Data Mining Society, 2021.

[133] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining Knowl. Discov.*, vol. 10, no. 3, 2020.

[134] U. Dogan, "Student engagement, academic self-efficacy, and academic motivation as predictors of academic performance," *The Anthropologist*, vol. 20, no. 3, pp. 553–561, 2015.

[135] J. Jovanovic, N. Mirriahi, D. Gasevic, S. Dawson, and A. Pardo, "Predictive power of regularity of pre-class activities in a flipped classroom," *Comput. Educ.*, vol. 134, pp. 156–168, 2019.

[136] M. S. Boroujeni, K. Sharma, L. Kidzinski, L. Lucignano, and P. Dillenbourg, "How to quantify student's regularity?," in *Proceedings of the 11th European Conference on Technology Enhanced Learning, EC-TEL 2016*, vol. 9891, pp. 277–291, Springer, 2016.

[137] S. Živkovil, "A model of critical thinking as an important attribute for success in the 21st century," *Procedia-social and behavioral sciences*, vol. 232, pp. 102–108, 2016.

[138] K. Hrbáčková, J. Hladík, and S. Vávrová, "The relationship between locus of control, metacognition, and academic success," *Procedia-Social and Behavioral Sciences*, vol. 69, pp. 1805–1811, 2012.

[139] D. Narang and S. Saini, "Metacognition and academic performance of rural adolescents," *Studies on home and community science*, vol. 7, no. 3, pp. 167–175, 2013.

[140] C. Berger, L. Alcalay, A. Torretti, and N. Milicic, "Socio-emotional well-being and academic achievement: Evidence from a multilevel approach," *Psicologia: reflexao e critica*, vol. 24, pp. 344–351, 2011.

[141] M. D. Toscano-Hermoso, C. Ruiz-Frutos, J. Fagundo-Rivera, J. Gómez-Salgado, J. J. García-Iglesias, and M. Romero-Martín, "Emotional intelligence and its relationship with emotional well-being and academic performance," *Children*, vol. 7, no. 12, p. 310, 2020.

[142] A. Rubio-Fernández, P. J. Muñoz-Merino, and C. D. Kloos, "A learning analytics tool for the support of the flipped classroom," *Comput. Appl. Eng. Educ.*, vol. 27, no. 5, pp. 1168–1185, 2019.

[143] S. M., S. S., S. Chatterjee, and K. Bijlani, "Learning analytics to identify students at-risk in moocs," in *Proceedings of the IEEE International Conference on Technology for Education*, pp. 194–199, 2016.

[144] T. Nazaretsky, M. Ariely, M. Cukurova, and G. Alexandron, "Teachers' trust in ai-powered educational technology and a professional development program to improve it," *Br. J. Educ. Technol.*, vol. 53, no. 4, pp. 914–931, 2022.

[145] Y. Tsai and D. Gasevic, "Learning analytics in higher education - challenges and policies: a review of eight learning analytics policies," in *Proceedings of the International Learning Analytics & Knowledge Conference*, pp. 233–242, ACM, 2017.

[146] T. Nazaretsky, M. Cukurova, and G. Alexandron, "An instrument for measuring teachers' trust in ai-based educational technology," in *Proceedings of the International Learning Analytics & Knowledge Conference*, pp. 56–66, 2022.

[147] J. Attenberg, P. Ipeirotis, and F. J. Provost, "Beat the machine: Challenging humans to find a predictive model's "unknown unknowns"," *ACM J. Data Inf. Qual.*, vol. 6, no. 1, pp. 1:1–1:17, 2015.

[148] G. Shabat, D. Segev, and A. Averbuch, "Uncovering unknown unknowns in financial services big data by unsupervised methodologies: Present and future trends," in *Proceedings of the KDD 2017 Workshop on Anomaly Detection in Finance*, vol. 71, pp. 8–19, PMLR, 2017.

[149] P. Zhao, Y. Zhang, and Z. Zhou, "Exploratory machine learning with unknown unknowns," in *Proceedings of the Symposium on Educational Advances in Artificial Intelligence*, pp. 10999–11006, AAAI Press, 2021.

[150] A. Liu, S. Guerra, I. Fung, G. Matute, E. Kamar, and W. Lasecki, "Towards hybrid human-ai workflows for unknown unknown detection," in *Proceedings of The Web Conference 2020*, pp. 2432–2442, 2020.

[151] V. Swamy, M. Marras, and T. Käser, "Meta transfer learning for early success prediction in moocs," in *Proceedings of the ACM Conference on Learning @ Scale*, pp. 121–132, ACM, 2022.

[152] C. Hardebolle, H. Verma, R. Tormey, and S. Deparis, "Gender, prior knowledge, and the impact of a flipped linear algebra course for engineers over multiple years," *Journal of Engineering Education*, 2022.

[153] J. Gardner and C. Brooks, "Student success prediction in moocs," *User Model. User Adapt. Interact.*, vol. 28, no. 2, pp. 127–203, 2018.

[154] S. V. Goidsenhoven, D. Bogdanova, G. Deeva, S. vanden Broucke, J. D. Weerdt, and M. Snoeck, "Predicting student success in a blended learning environment," in *Proceedings of the International Learning Analytics & Knowledge Conference*, pp. 17–25, ACM, 2020.

[155] V. Swamy, B. Radmehr, N. Krco, M. Marras, and T. Käser, "Evaluating the explainers: Black-box explainable machine learning for student success prediction in moocs," in *Proceedings of the International Conference on Educational Data Mining*, 2022.

[156] E. B. Nilsen, D. E. Bowler, and J. D. Linnell, "Exploratory and confirmatory research in the open science era," *Journal of Applied Ecology*, vol. 57, no. 4, pp. 842–847, 2020.

[157] Y. Zheng, N. Ghane, and M. Sabouri, "Personalized educational learning with multi-stakeholder optimizations," in *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pp. 283–289, 2019.

[158] M. Cevikbas and G. Kaiser, "Promoting personalized learning in flipped classrooms: A systematic review study," *Sustainability*, vol. 14, no. 18, p. 11393, 2022.

[159] M. Sahin and D. Ifenthaler, "Visualizations and dashboards for learning analytics: A systematic literature review," *Visualizations and dashboards for learning analytics*, pp. 3–22, 2021.

[160] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, pp. 1–21, 2020.

[161] I. Jivet, M. Scheffel, H. Drachsler, and M. Specht, "Awareness is not enough: Pitfalls of learning analytics dashboards in the educational practice," in *Proc. European Conference on Technology Enhanced Learning*, pp. 82–96, Springer, 2017.

[162] J. Attenberg, P. Ipeirotis, and F. Provost, "Beat the machine: Challenging humans to find a predictive model's "unknown unknowns"," *Journal of Data and Information Quality (JDIQ)*, vol. 6, no. 1, pp. 1–17, 2015.

[163] P. Zhao, Y.-J. Zhang, and Z.-H. Zhou, "Exploratory machine learning with unknown unknowns," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 10999–11006, 2021.

[164] G. Shabat, D. Segev, and A. Averbuch, "Uncovering unknown unknowns in financial services big data by unsupervised methodologies: Present and future trends," in *KDD 2017 Workshop on Anomaly Detection in Finance*, pp. 8–19, PMLR, 2018.

[165] V. Sher, M. Hatala, and D. Gašević, "Analyzing the consistency in within-activity learning patterns in blended learning," in *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pp. 1–10, 2020.

[166] M. Marras, J. T. T. Vignoud, and T. Kaser, "Can feature predictive power generalize? benchmarking early predictors of student success across flipped and online courses," in *14th International Conference on Educational Data Mining*, pp. 150–160, 2021.

[167] V. Swamy, M. Marras, and T. Käser, "Meta transfer learning for early success prediction in moocs," in *Proceedings of the Ninth ACM Conference on Learning@ Scale*, pp. 121–132, 2022.

[168] K. Verbert, S. Govaerts, E. Duval, J. L. Santos, F. Van Assche, G. Parra, and J. Klerkx, "Learning dashboards: an overview and future research opportunities," *Personal and Ubiquitous Computing*, vol. 18, pp. 1499–1514, 2014.

[169] R. Liu, J. An, Z. Wang, J. Guan, J. Liu, J. Jiang, Z. Chen, H. Li, B. Peng, and X. Wang, "Artificial intelligence in laparoscopic cholecystectomy: does computer vision outperform human vision?," *Artificial Intelligence Surgery*, vol. 2, no. 2, pp. 80–92, 2022.

[170] R. Rajalingham, E. B. Issa, P. Bashivan, K. Kar, K. Schmidt, and J. J. DiCarlo, "Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks," *Journal of Neuroscience*, vol. 38, no. 33, pp. 7255–7269, 2018.

[171] Z. Zhang, J. Singh, U. Gadiraju, and A. Anand, "Dissonance between human and machine understanding," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–23, 2019.

[172] A. Boldt, V. De Gardelle, and N. Yeung, "The impact of evidence reliability on sensitivity and bias in decision confidence.," *Journal of experimental psychology: human perception and performance*, vol. 43, no. 8, p. 1520, 2017.

[173] M. Beardsley, L. Albó, P. Aragón, and D. Hernández-Leo, "Emergency education effects on teacher abilities and motivation to use digital technologies," *British Journal of Educational Technology*, vol. 52, no. 4, pp. 1455–1477, 2021.

[174] A. Pardo and G. Siemens, "Ethical and privacy principles for learning analytics," *British journal of educational technology*, vol. 45, no. 3, pp. 438–450, 2014.