Towards Transparent and Trustworthy Prediction of Student Learning Achievement by Including Instructors as Co-Designers: A Case Study

Xiaojing Duan[1], Pei Bo[1], G. Alex Ambrose[1], Arnon Hershkovitz[2], Ying Cheng[1], Chaoli Wang[1]

[1]University of Notre Dame, Notre Dame, IN, USA
[2] Tel Aviv University, P.O.Box 39040, Tel-Aviv, 699780, IL, Israel

Providing educators with understandable, actionable, and trustworthy insights drawn from large-scope heterogeneous learning data is of paramount importance in achieving the full potential of artificial intelligence (AI) in educational settings. Explainable AI (XAI)—contrary to the traditional "black-box" approach—helps fulfilling this important goal. We present a case study of building prediction models for undergraduate students' learning achievement in a Computer Science course, where the development process involves the course instructor as a co-designer, and with the use of XAI technologies to explain the underlying reasoning of several machine learning predictions. The explanations enhance the transparency of the predictions and open the door for educators to share their judgments and insights. It further enables us to refine the predictions by incorporating the educators' contextual knowledge of the course and of the students. Through this human-AI collaboration process, we demonstrate how to achieve a more accountable understanding of students' learning and drive towards transparent and trustworthy student learning achievement prediction by keeping instructors in the loop. Our study highlights that trustworthy AI in education should emphasize not only the interpretability of the predicted outcomes and prediction process, but also the incorporation of subject-matter experts throughout the development of prediction models.

Additional Key Words and Phrases: Student learning achievement prediction, Transparent and trustworthy AI, Explainable AI (XAI), Human-centered AI, Learning analytics, Co-design

1 INTRODUCTION

In the transition to a digital era, artificial intelligence (AI) has been applied in divergent applications in education. These applications include examining the student learning process for providing personalized and timely interventions in support of online learning practices (Kim, 2020; Zhang, 2020), modeling student learning process for the identification of at-risk students who might drop out of the courses at an early stage (Goel, 2020; Kloft, 2014), and tracing student historical learning paths for providing individualized learning recommendations (Takami , 2022; Heras, 2020; Ndiyae, 2019). Practically, other applications, like AI-powered augmented and virtual reality (AR/VR) (Rong, 2022), wearable devices (Ciolacu, 2021), speech-to-text and text-to-speech applications (Azeta, 2009), etc. have also been implemented to actually help students with either hearing or visual impairments. Despite the effectiveness as well as the potential benefits of AI in education, most of these applications are still at a nascent, experimental stage rather than a systematic level (Vincent-Lancrin, 2020; Szafir, 2013). While the reasons limiting the use of AI education can be various, studies have shown that the trustworthiness of the results from AI algorithms and the transparency of the prediction process, especially in sometimes high-stake situations, have been the major factors that undermine the realization of the full potential of AI in educational settings (Anwar, 2021).

Understanding the behaviors of AI, as well as promoting the trustworthiness and transparency of AI in education, becomes more important, especially since education is undergoing a transformation currently with the increasing incorporation of digital technologies. Specifically, in education, trustworthy and transparent AI has multiple dimensions. On the one hand, AI might be considered trustworthy when it does what it is supposed to do and generates insights that align

with human values. For example, AI-powered learning analytic systems identify at-risk students through profiling student learning patterns based on recorded learning activities in the online learning systems. If the effectiveness of the algorithms is limited by recognizing students who spent less time on assignments as the ones who might drop out of the course, merely relying on this insight will mislead the future teaching and instructional practices in the large population considering struggling students who spent more time on assignments as proficient ones. On the other hand, trustworthy AI algorithms characterize student learning processes accurately and, at the same time, provide interpretations of the results for instructors. Currently, inconsistencies have been identified by multiple research in terms of the purposes of AI algorithm designers and the requirements of the actual situations (Conati, 2018; Mahbooba, 2021). Most current AI algorithms focus on the accuracy of predictions, which always leads to sometimes complicated and unexplainable black-box models. However, these models seem meaningless for practices in educational settings that require actionable insights to really help students' learning rather than accurately predict which students will fail in the end. These dimensions suggest that it is necessary to unpack the black-box of educational AI applications and incorporate them with subjective contextual knowledge as well as human perceptions to make them more aligned with the needs in real settings.

The Explainable AI (XAI) is gaining popularity due to the capabilities of automatically providing interpretations about how the predictions were achieved and which factors had the most significant correlation with the predictions based on the built models. This provides an important interface between the abstract and complex AI algorithms with the end-users (e.g., learners, educators, administrators) who are domain experts but might have no or limited background knowledge regarding AI technologies. In fact, researchers have explored various

approaches for generating interpretable and understandable predictions based on the architecture of current AI systems. At the very early stage, the explainable techniques are mainly those AI model-specific, such as Decision Tree, Linear Regression, Bayesian Networks, and so on. More recently, model-agnostic approaches have been proposed for generalizing interpretation techniques to all machine learning models. SHapley Additive exPlanations (SHAP) is such a model-agnostic approach (Lundberg, 2017). It uses the classic Shapley values from game theory and their related extensions to explain the contributions of each feature to model predictions (Li, 2022). Compared to other explainable approaches, the useful properties such as efficiency, summary, and additivity make the SHAP framework appealing to various applications from financial to medical (Lu, 2021; Mokhtari, 2019).

Our case study explores the potential of using the SHAP framework, while involving instructors' contextual insights in the development of a prediction model, to promote trustworthy and transparent predictions of students' learning achievement. Unlike most of the existing student success prediction practices in which students are classified into dichotomy classes (i.e., Pass or Failure), our study employs different regression machine learning models to predict students' performance in the final exam because it depicts a more complete picture of students' learning. Moreover, we employ the Kernel SHAP method to explain how the models derive the predictions by examining the contribution and impact of each feature on the prediction outcomes. In doing so, our study has two contributions: (1) We demonstrate how to enhance the prediction's transparency by explaining the underlying reasoning of various machine learning models. The explanation allows the end-users of the models to judge the prediction outcomes and share their contextual insights. (2) We explicitly incorporate the instructor as a co-designer of the prediction models, therefore incorporating his contextual knowledge as well as his

perceptions of the reasoning behind the prediction. As a result, our case study presents the initial step toward transparent and trustworthy prediction of student learning achievement by combining the XAI technologies and the instructor's contextual insights.

2 RELATED WORK

With the increasing applications of AI in various fields, trust has become a central component in Human and AI interactions. Trustworthy interactions can promote the incorporation of AI into society in a safer manner (Shneiderman, 2020). It is often interlinked with interpretations and quality of results and is associated with the inclusion of supportive information to boost trustworthiness. Currently, most applications focus on building high-accuracy predictive AI models to accurately reflect student learning from large heterogeneous data, with few considering the trustworthiness of these models (Murdoch, 2019). Therefore, improving the trustworthiness of AI, especially in high-stake contexts, has become a major focus of the most recent research.

Promoting trustworthiness in using AI can be beneficial in facilitating and sustaining collaborative relationships between end-users and AI developers to generate more informed decision-making processes (Thornton, 2021). For example, collective insights about student learning status rely not only on the analytical results from large-scale learning data using some AI models but also on instructors' domain knowledge about the difficulties of the current learning contents as well as the perceptions about students' historical learning status. Without a mechanism to facilitate educators with different knowledge backgrounds about AI to trust the predictions from the seemingly "black-box" AI algorithms, it is difficult to inject the contextual knowledge from instructors into the objective predictions from models to generate a comprehensive understanding of students' learning. Additionally, there is also a huge potential

that the incorrect levels of trust in the predictions could lead to misuse, abuse, and disuse of AI technology in making decisions in some high-stake situations (Jacovi, 2021). Floridi et al. (2018) explored non-maleficence principles to promote the development, deployment, and use of AI for human well-being. By investigating core opportunities and risks of AI in society as well as identifying several ethical considerations that should undergird the development and adoption of AI, the principle specifically concerns the aspects of human privacy, security, and safety.

Achieving trust in human-AI collaboration in education can be challenging (Nazaretsky, 2022). On the one hand, most of the current AI-powered models are always built on highly complex, multi-scale, and interconnected environmental data to generate higher accuracy predictive models, making it difficult to identify the aspects of trust and further explore assessment metrics (Thornton, 2021). Particularly in education settings, teaching and learning practices are often complex processes involving multiple, nonlinear interactions between instructional strategies and different learning dimensions measured in terms of both cognitive and physical aspects. Consequently, there are variegated assemblage learning data in terms of structured or unstructured data at different levels of variety and veracity. As such, much research (Al-Shabandar, 2019; Kurdi, 2020) mainly focuses on extracting learning patterns and building automatic models from these data to predict students' performance, which leads to high-complexity AI models with less interpretability. On the other hand, there is currently limited understanding of trust interactions with AI, given the multiple dimensions in both individual and group interactions (Jacovi, 2021). Considering the interdisciplinary and dynamic properties of Human-AI interactions (Thiebes, 2021), previous research characterized the trust of interactions from both individual and public levels. Jacovi et al. (2021), focusing on individual interactions with AI, defined the major properties of trust as the vulnerability of the user and the ability to

anticipate the impact of models' decisions. The authors explicitly discussed the conditions under which trust occurs and the prerequisites that trust goals could be achieved. Additionally, Knowles and Richards (2019) further provided an overview of the theoretical framework to foster public trust in AI. The authors specifically distinguished the expert trust and public trust in AI by correcting the general misunderstandings of trust in AI.

These studies provided the definitions of trustworthiness and transparency as well as explored the characteristics of technologies that promote accountable use of AI for a general purpose (Jacovi, 2021). However, few of them explicitly explore trustworthy AI situated in educational settings and examine how trustworthiness can be used to better support real teaching and learning practices. Currently, XAI has been widely used to justify the reliability and trustworthiness of AI algorithms based on the data in a particular context. Mahbooba et al. (2021) employed XAI in combination with the Decision Tree (DT) model to support human experts in understanding malicious data and detecting the intrusions of a system by providing the reasoning process of the DT model used in the intrusion detection context. Moreover, Khosravi et al. (2022) and Swamy et al. (2022) explicitly investigated XAI in educational settings. Specifically, an XAI-ED framework was developed by Khosravi et al. (2022) to support the explainability of AI models applied in education as well as present the benefits and pitfalls of providing explanations within such context. Swamy et al. (2022) compared the explanatory results of 5 different XAI models (i.e., LIME, PermutatsionSHAP, KernelSHAP, DiCE, CEM) in the situation of applying Deep learning approaches to predict student learning. The results suggest that the explanations of results can be associated with the choice of explanatory models. These studies provided interfaces and proxies for end-users to understand the underlying causal reasoning of AI models in a specific context. Few of them further explored the real impacts of

the explanations on end-users and how the new insights generated from end-users can be used to refine AI models. Indeed, involving stakeholders as co-designers of learning analytics is still in its infancy (Sarmiento & Wise, 2022). To bridge these gaps, our study first explores the potential of utilizing the SHAP framework to explain the reasoning behind various machine learning models. We then investigate how to leverage the instructor's contextual insights to refine the predictions, make them more trustworthy, and gain a more accountable understanding of student learning. Formally, we attempt to address the following two research questions:

- RQ1. How do we enhance the transparency of black-box predictive models by explaining their predictions?
- RQ2. How do we incorporate instructors' contextual knowledge to improve the trustworthiness of predictions and generate more plausible insights about student learning?

3 METHOD

3.1 Study Context and Data Collection

The case study was conducted on a Fundamentals of Computing course offered in Fall 2021 at a midwestern university in the United States. The course lasted for 15 weeks and had 131 students enrolled in it. The course is required for all computer science and computer engineering students. It aims to help students develop basic proficiency in the C programming language and formulate algorithms to solve computational problems in different domains. The course consisted of 11 labs, nine exercise assignments, and three exams. The labs were submitted electronically, and for each lab, we collected the timestamp of students' last submission attempts. The lab attendance was graded. The exercise assignments were conducted in zyBooks, an interactive digital textbooks platform, and we collected the time duration that students spent on those exercises.

The first and second exams were conducted during the 7th and 11th weeks of the course, and the final exam was conducted in the last week, i.e., the 15th week.

3.2 Feature Extraction and Model Development

One of our goals was to identify the potentially low-performing students. Many previous studies used the course pass or failure status as an indicator of students' learning achievement and built machine learning models to predict the binary label (Li, 2022; Lee, 2015; Macfadyen, 2010; Romero, 2013; Syed, 2019). We found that approach has limitations because the binary label does not completely depict a student's mastery of the course content. After discussing with the instructor, we decided to predict students' performance in the final exam (FinalExamScore) because it assesses students' mastery of all the topics covered in the course and therefore is a more comprehensive indicator of their learning achievement. To predict students' FinalExamScore, we first extracted a set of students' performance and learning behaviors features that are commonly used to predict students' learning achievement in the literature (Rotelli , 2022; Duan, 2022; Marras, 2021; Matcha, 2019). We normalized each feature using the MinMax scaling so all the features are on the same scale. Then we split the data randomly into training and testing sets to train four machine learning models, including LinearRegression, RandomForestRegressor, StochasticGradientDescentRegressor (SGDRegressor), and SupportVectorMachineRegressor (SVR). The ratios for the training and testing sets are 70% and 30%, respectively. After performing hyperparameter tuning on the models and evaluating their performance on the testing set with different combinations of the features, we settled on the list of features described in Table 1 because they resulted in the best performance in terms of the lowest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Those features reflect both students' learning behavior and course performance.

Table 1. Features Name and Description

Name	Description	Category
Exam1Score	A student's Exam1 score	Course performance
Exam2Score	A student's Exam2 score	Course performance
ExerciseScore	The sum of a student's 9 exercise assignment scores	Course performance
ExerciseTime	The total time in minutes a student spent on the exercise assignment	Learning behavior
LabScore	The sum of a student's 11 lab scores	Course performance
LabAttendanceScore	The sum of a student's 10 lab attendance scores	Learning behavior

The resulting MAE and RMSE of each model are described in Table 2. As it shows, LinearRegression has the lowest MAE and RMSE rates.

Table 2. Model Performance Metrics

	LinearRegression	RandomForestRegressor	SGDRegressor	SVR
MAE	7.74	8.37	8.22	9.96
RMSE	12.17	12.51	12.58	16.15

3.3 Model Explanation

To enhance the model's transparency, we used the Kernel SHAP method (Lundberg, 2017) to explain how the model made the prediction at both the global and local levels. Kernel SHAP uses a special weighted linear regression to compute the Shapley values based on the coalitional game theory. The Shapley values calculated using the conditional expectations are called SHAP (SHapley Additive exPlanation) values. They describe feature importance with the consideration of different feature subsets and their effects on the model's predictions. A feature's SHAP value represents its significance in the model prediction and is computed via formula (1):

$$\emptyset_i = \sum_{S \subseteq F\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}\left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)\right] \quad (1)$$

where $\emptyset_i$ is the SHAP value for feature i, S is a subset of all features F, and x_S represents the values of the input features in the subset S. A model $f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right)$ is trained with the feature i

and another model $f_S(x_S)$ is trained without the feature i. $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$ is the difference of the two models' predictions on the current input. It is then computed for all the possible subsets $S \subseteq F\{i\}$ because the effect of withholding a feature depends on other features. The SHAP value $\emptyset_i$, i.e., feature i's significance, is a weighted average of all the possible differences. This approach has become the preferred method for model-agnostic SHAP value calculation (Lundberg, 2017).

At the global level, we showed the features' overall significance and their impact on the prediction. A feature's overall significance is calculated by averaging its absolute SHAP value in each prediction. A feature's impact is derived by examining its contribution to the prediction given its value. The higher the value, the greater the impact. At the local level, we explained how a model predicted for an individual student by showing each feature's SHAP value. The positive SHAP value indicates the feature contributes positively to the prediction, while a negative SHAP value indicates it contributes negatively to the prediction. A SHAP value of zero indicates the feature has little to no significance to the prediction.

4 RESULTS

4.1 Enhancing Prediction's Transparency

To address our first research question, i.e., enhancing the transparency of the prediction, we uncovered the features' overall significance and their impacts on the prediction. We also revealed how the individual prediction was derived.

4.1.1 Uncovering the features' overall significance and their impacts on the prediction

The overall significance of each feature in the prediction is described in Fig. 1. As it shows, the LinearRegression and RandomForestRegressor models agree that Exam2Score is the most

significant feature for predicting FinalExamScore. However, SGDRegressor and SVR disagree with that. The most significant feature in SGDRegressor and SVR are the LabScore and Exam1Score, respectively. For the least significant feature, LinearRegression and SGDRegressor agree on ExerciseTime, while RandomForestRegressor and SVR points to LabAttendanceScore and ExerciseScore, respectively.

Feature	LinearRegression	RandomForestRegressor	SGDRegressor	SVR
Exam1Score	2.45	2.68	2.53	1.67
Exam2Score	4.12	5.62	2.43	0.50
ExerciseScore	1.35	1.33	0.68	0.14
ExerciseTime	0.14	0.61	0.64	0.74
LabAttendanceScore	1.14	0.27	0.94	0.14
LabScore	2.79	3.09	3.19	0.65

Fig. 1. Features' significance to the prediction. The most and least significant values are depicted in red and blue color, respectively.

To uncover the impact of each feature on the prediction, we visualized the relationship between the feature value and its contribution to the prediction, as illustrated in Fig. 2. The color of a dot denotes the value of its represented feature. The bluer the color, the lower the value. The redder the color, the higher the value. The horizontal distance between a dot and the vertical zero-line depicts the contribution to the prediction made by the feature value represented by the dot. Because there are many instances of duplicated feature values, we jittered the graph to avoid clutter. As Fig. 2 shows, all the models agree that Exam1Score, Exam2Score, LabAttendanceScore, and LabScore have a positive impact on the prediction, meaning that the higher the value of those features, the higher the predicted FinalExamScore. However, the models disagree on the impact of ExercisesScore and ExerciseTime. ExercisesScore negatively impacts the prediction in LinearRegression, RandomForestRegressor, and SGDRegressorut it

exhibits a positive impact on the SVR's prediction. ExerciseTime negatively impacts the prediction in RandomForestRegressor, SGDRegressor, and SVR, but it exhibits a positive impact on the LinearRegression's prediction.

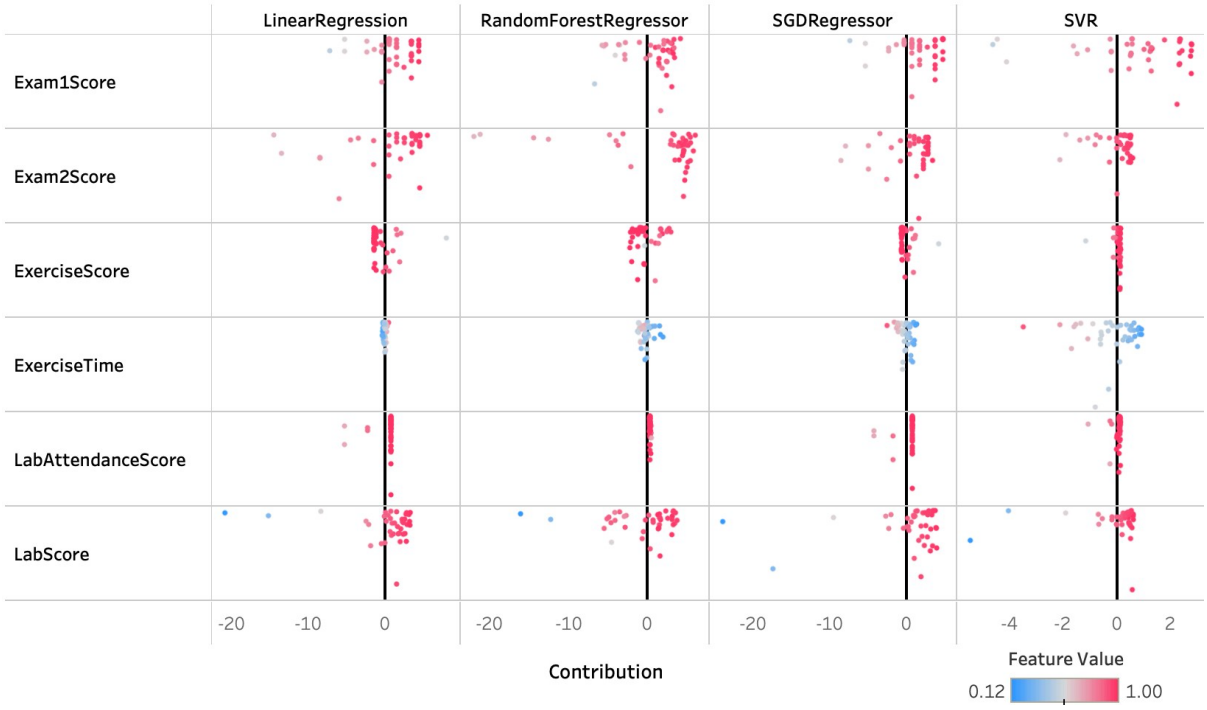

Fig. 2. Features' impact on the prediction in each model.

To further examine this observation, we visualized the relationship between the ExerciseScore and ExerciseTime values and their contributions to the prediction in scatterplots, as shown in Fig. 3. Those plots confirm our observations in Fig. 2.

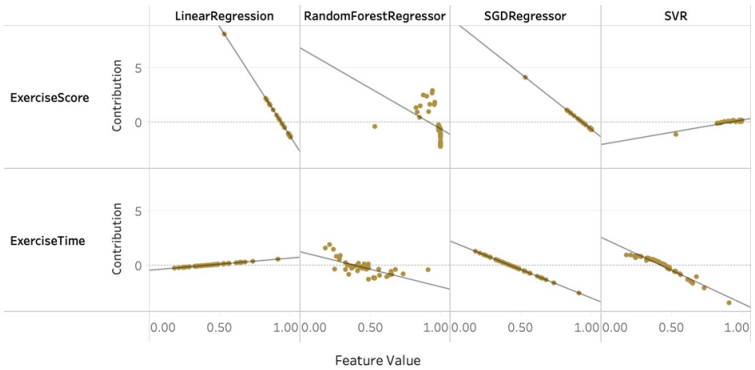

Fig. 3 ExerciseScore vs its contribution and ExerciseTime vs its contribution to the prediction.

After sharing these results with the instructor, he found the negative impact of ExerciseTime on the prediction acceptable. He commented it is not uncommon that high-performing students can complete the exercises in less time than low-performing students. However, he was puzzled by the negative impact of ExerciseScore on the prediction demonstrated in three out of four models. His previous experience has shown that a higher exercise score indicates better mastery of the course content. Therefore, it is counterintuitive that the higher exercise score contributes less to the prediction and the lower exercise score contributes more to the prediction. This feedback prompted us to investigate the prediction for individual students.

4.1.2 Revealing how the prediction for individual students was made

We revealed how each model made the prediction for individual students by explaining the contribution of each feature to that particular prediction, that is, at the instance level. In this section, we illustrate how the model works at the instance level, that is, for individual students. This will be demonstrated on two random students that were chosen from the testing set, namely, Student1 and Student7. Table 3 describes their true FinalExamScore and predicted scores. For Student1, LinearRegression's prediction (86.18) is the closest to the true score (85). For Student7, RandomForestRegressor's prediction (80.79) is the closest to the true score (82).

Table 3. Student1 and Student7's True and Predicted FinalExamScore

	TrueScore	LinearRegression	RandomForestRegressor	SGDRegressor	SVR
Student1	85	86.18	83.14	86.40	86.20
Student7	82	88.35	80.79	87.76	86.07

Fig. 4 illustrates each feature's contribution to the predicted FinalExamScore for Student1 and Student7 in each model. The green Gantt bars mark the value of each feature. The horizontal bars represent the contribution of each feature to the prediction, with the red color denoting positive

contribution and the blue color denoting negative contribution. For example, Student1's ExerciseScore (1) contributed -1.36, -2.19, -0.69, and 0.13 to the LinearRegression, RandomForestRegressor, SGDRegressor, and SVR's prediction, respectively. This student's ExerciseTime (0.371) contributed -0.06, -0.29, 0.29, and 0.50 to the LinearRegression, RandomForestRegressor, SGDRegressor, and SVR's prediction, respectively. Student 7's ExerciseScore (0.831) contributed 2.17, 1.32, 1.10, and -0.12 to the LinearRegression, RandomForestRegressor, SGDRegressor, and SVR's prediction, respectively. This student's ExerciseTime (0.491) contributed 0.07, -1.31, -0.34, and -0.23 to the LinearRegression, RandomForestRegressor, SGDRegressor, and SVR's prediction, respectively.

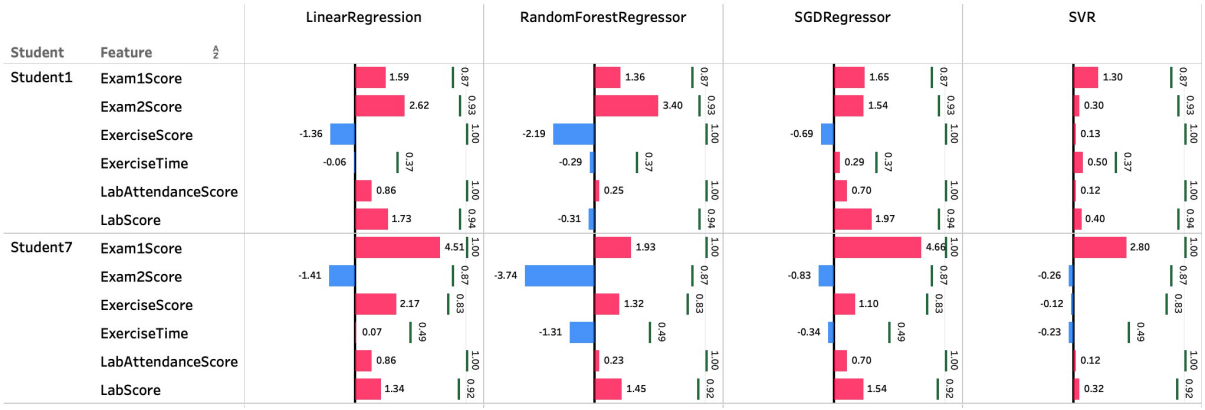

Fig. 4. Feature contribution to the predicted FinalExamScore for Student1 and Student7.

Fig. 5 compares the contributions of ExerciseTime and ExerciseScore to the predicted FinalExamScore for Student1 and Student7. As it shows, Student7's lower ExerciseScore (0.831) had a more positive contribution to the prediction than Student1's higher ExerciseScore (11) did in all the models except for SVR. Additionally, Student7's higher ExerciseTime (0.491) had a more negative contribution to the prediction than Student1's lower ExerciseTime (0.371) did in all the models except for LinearRegression. These results are consistent with the overall negative impact of ExerciseTime and ExerciseScore on the prediction described in the previous section.

After sharing them with the instructor, he did not voice concern over the negative impact of ExerciseTime on the prediction. However, he could not be convinced that ExerciseScore hurts students' final exam performance. He was concerned about the potential harm it can cause students when presenting them with a lower predicted FinalExamScore given they have higher ExerciseScore. Therefore, he suspected that those models could be used to identify potential low-performing students in real-world settings. His concern and suspicion drove us to pursue the second research question.

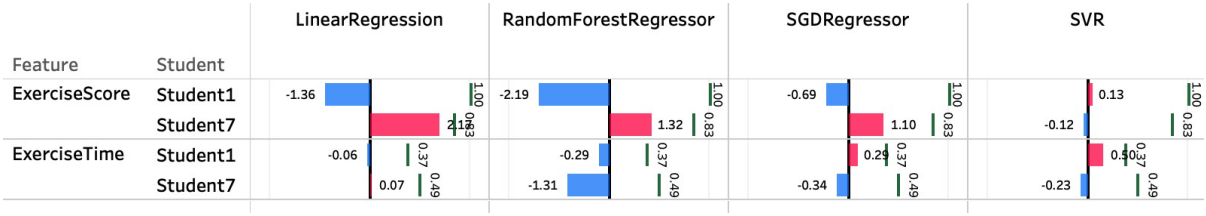

Fig. 5. Comparison of ExerciseTime and ExerciseScore's contribution to the predicted FinalExamScore for Student1 and Student7.

4.2 Improve Prediction's Trustworthiness

To address our second research question, i.e., improving the trustworthiness of the prediction, we first investigated the cause of the negative impact of ExerciseScore on the prediction and then took proper actions to mitigate it.

4.2.1 Investigating the cause of the negative impact

We grouped students by their FinalExamScore, ranked the groups, and compared the average and median ExerciseScore and ExercisesTime between the groups. As the results show in Table 4, the average ExerciseScore of both the very high and high FinalExamScore groups is lower than that of the fair FinalExamScore group. Likewise, the average ExerciseScore of both the high and acceptable groups is lower than that of the low FinalExamScore group. Formally, ExcerciseScore and FinalExamScore do not exhibit a positive correlation. This data bias explains

the negative contribution of ExerciseScore to the prediction in three of our four models. Table 4 also shows that the higher ranked a group's FinalExamScore, the less time the group spent on the exercises on average. The only exception is the very low FinalExamScore group which spent the least amount of time on average. This confirms the instructor's hypothesis that the high-performing students can complete the exercises in less time than the low-performing group.

Table 4. Average ExerciseScore and ExerciseTime of Students Grouped by FinalExamScore

FinalExamScore Group	Student Count	ExerciseScore Mean (std)	ExerciseTime Mean (std)
50 or below (very low)	5	293 (74.11)	200 (99.64)
51 - 60 (low)	6	380.33 (17.95)	445.50 (175.97)
61 - 70 (acceptable)	16	357.38 (59.49)	367.88 (135.24)
71 - 80 (fair)	22	388 (14.61)	359.95 (88.32)
81 - 90 (high)	42	373.21 (63.29)	331.64 (121.82)
91 - 100 (very high)	40	382.38 (28.73)	294.95 (95.75)

4.2.2 Mitigating the data bias and improving prediction's trustworthiness.

We first log-scaled the ExerciseScore and followed our previous process to train the predictive models. The log scale reduced the overall impact of ExerciseScore on the prediction but did not change its negative contribution. Next, together with the instructor, we re-evaluated our features, aiming at finding features that could better reflect student learning achievement. His domain knowledge and teaching experience led us to generate a new feature by dividing ExerciseScore by ExerciseTime. This new feature measures students' learning gains while considering the time they dedicated to the exercies and may better reflect their effort or efficacy. We named the new feature ExerciseScorePerMin and replaced ExerciseScore and ExerciseTime with it to train the LinerRegression, RandomForestRegression, SGDRegressor, and SVR models. The performance metrics are described in Table 5. Compared to the metrics resulting from using the ExerciseScore and ExerciseTime features, we can see the new ExerciseScorePerMin feature improves the performance of LinearRegression, RandomForestRegressor, and SGDRegressor. In more detail,

the LinearRegression's RMSE decreased from 12.17 to 11.88. The RandomForestRegressor's RMSE decreased from 12.51 to 11.79. The SGDRegressor's RMSE decreased from 12.58 to 12.43. SVR is the only exception, and its RMSE increased slightly from 16.15 to 16.26. The new result also suggests that RandomForestRegressor has the best performance.

Table 5. Model Performance Metrics After Replacing ExerciseScore and ExerciseTime with ExerciseScorePerMin

	LinearRegression	RandomForestRegressor	SGDRegressor	SVR
MAE	7.54	7.50	8.03	9.99
RMSE	11.88	11.79	12.43	16.26

We also examined how the change of feature influenced the overall features' significance and their impact on the prediction. As the results show in Fig. 6, the most important feature stays the same in all models as it was before the transformation (compare with Fig. 1). The least important feature of LinearRegression and SGDRegressor becomes ExerciseScorePerMin. In RandomForestRegressor, LabAttendanceScore remains the least important feature. The SVR model agrees with that now.

Feature	LinearRegression	RandomForestRegressor	SGDRegressor	SVR
Exam1Score	2.44	2.79	2.62	1.61
Exam2Score	4.36	4.91	2.37	0.49
ExerciseScorePerMin	0.13	0.59	0.13	1.00
LabAttendanceScore	0.52	0.36	0.68	0.09
LabScore	2.51	2.90	2.97	0.64

Fig. 6. Features' significance to the prediction after replacing ExerciseScore and ExerciseTime with ExerciseScorePerMin.

The features' impact on prediction after replacing ExerciseScore and ExerciseTime with ExerciseScorePerMin is illustrated in Fig. 7. As it shows, all the models agree that all the features now positively impact the prediction.

Fig. 7. Features' impact on the prediction after replacing ExerciseScore and ExerciseTime with ExerciseScorePerMin.

We further examined the impact of the new ExerciseScorePerMin feature on the prediction. As shown in Fig. 8, it is positive in all the models now.

Fig. 8. Relationship between ExerciseScorePerMin value and its contributions to the prediction.

Additionally, we investigated how the new feature influenced the models' prediction for individual students. We continued to use Student1 and Student7 as the testbed. Table 6 describes

the new prediction of their FinalExamScore. For Student1, RandomForestRegressor's prediction (85.31) is the closest to the true score (85). It is an improvement compared to the best prediction (86.18) resulting from using the ExerciseScore and ExerciseTime features. For Student7, RandomForestRegressor's prediction (81.78) is the closest to the true score (82). It is an improvement compared to the best prediction (80.79) due to using the ExerciseScore and ExerciseTime features.

Table 6. Student1 and Student7's True and Predicted FinalExamScore after Replacing ExerciseScore and ExerciseTime with ExerciseScorePerMin.

	True Score	LinearRegression	RandomForestRegressor	SGDRegressor	SVR
Student1	85	87.118	85.31	86.55	85.88
Student7	82	85.16	81.78	86.69	84.89

Fig. 9 illustrates how the predictions for Student1 and Student7 were formed after replacing ExerciseScore and ExerciseTime with ExerciseScorePerMin. As it shows, Student1's ExerciseScorePerMin (0.362) contributed 0.04, 0.04, 0.04, and 0.58 to the predictions in LinearRegression, RandomForestRegressor, SGDRegressor, and SVR, respectively. Student7's ExerciseScorePerMin (0.15) contributed -0.14, -1.01, -0.14, and -1.17 to the predictions in LinearRegression, RandomForestRegressor, SGDRegressor, and SVR, respectively.

Fig. 9. Feature Contribution to the predicted FinalExamScore for Student1 and Student7 after replacing ExerciseScore and ExerciseTime with ExerciseScorePerMin.

Comparing ExerciseScorePerMin's contribution to the predictions for Student1 and Student7, it is obvious that the higher value contributed more to the prediction, and the lower value contributed less to the prediction in all the models. The instructor found this new result more intuitive. He commented that ExerciseScorePerMin was a better feature for predicting students' FinalExamScore, leading to more trustworthy predictions.

5 DISCUSSION

Machine learning models have been increasingly used in educational settings to identify potentially at-risk students and trigger proper interventions aimed to help them achieve academic success. However, the models' complex structures and hidden decision-making mechanisms made it difficult for end-users to interpret and trust their predictions. In this case study, we demonstrated how to enhance the transparency of various machine learning models (RQ1). More specifically, we applied the SHAP framework to uncover the features' overall significance and their impact on the prediction of students' learning achievement.

As shown in our results, the features' importance and their impact on the prediction vary across models. For example, students' grades on the course exercises exhibited a negative impact on the prediction in some regression models (LinearRegression, RandomForestRegressor, and SGDRegressor), but a positive impact in another (SVR); similarly, the time students dedicated to course exercises demonstrated a negative impact in some regression models (RandomForestRegressor, SGDRegressor, and SVR), but a positive impact in another (LinearRegression). The instructor found these results puzzling. Our study showcased how to leverage the instructor's contextual knowledge to solve the puzzle. It has been shown in previous studies that time dedicated to course exercises can be negatively associated with course grade (De Jong, 2000; Kitsantas, 2011; Trautwein, 2007). As the more time a student dedicates to

exercises, the less proficient, efficient, or motivated this student may be, which may impact negatively their final grades. However, exercise score's negative impact on the prediction was more counterintuitive, and not in line with the literature (Cheema, 2015, Fan, 2017; Fernández-Alonso, 2015), which had led us to discover a bias in the data. We mitigated this bias by conducting feature engineering through a co-design process with the instructor. His contextual knowledge informed us to generate a new feature which takes the ratio of students' learning gains in exercises over the time they dedicated, to better reflect effort or efficiency. It was previously shown that this variable still captures contextual behavior and is not fixed (Hershkovitz & Nachmias, 2009), and indeed it has been used to measure and study learning in various settings (e.g., Ben-Zadok, Leiba, & Nachmias, 2010; Hänsch et al., 2018; Pejić & Stanić Molcer,2021; Zhang, Guo, & Lius, 2021). The new feature was indeed found to have a positive impact on students' predicted learning achievement. Additionally, not only that this new feature improved the predictions of most of our models, but its relationship with the target variable was aligned with the instructor's teaching experience and domain knowledge. Therefore, the instructor found the predictions more trustworthy and became more comfortable with using it to identify the low-performing students (RQ2). This co-design process, which resulted in an improved, more trustworthy prediction model, demonstrates the power of expert feature engineering; indeed, it was suggested that when wishing to predict students' behavior in an easy-to-interpret way, expert feature engineering has an important role, and therefore it should be incorporated into machine learning-based model building (Botelho, Baker, & Heffernan, 2019; Levin, 2021; Jiang et al., 2018).

An important implication of our study is that AI has the potential to assist educators with their effort in course design improvement. In our case study, we observed that high-performing

students received lower exercise scores and spent less time on them. In comparison, low-performing students received higher exercise scores by spending more time on them. These findings open the opportunity for instructors to reassess the course and exercise design. The high-performing students might find the exercises less challenging and were not motivated to spend more effort on it. When encountering this situation, instructors could add more challenging exercises and allocating more points to the challenging questions to motivate the high-performing students. The low-performing students might have trouble applying the knowledge they have gained through the exercises to solve more advanced problems. When facing this challenge, instructors could tailor the lecture toward demoing how to synthesize the various topics and apply them to solve more comprehensive problems. Our study also highlights the importance of improving the alignment of course exercises and the final exam. While exercises typically focus on individual concepts, the final exam assesses students' comprehensive understanding of the course material. Therefore, some students who excel in individual concepts may struggle to combine them and perform poorly in the final exam. To better prepare students for the final exam, instructors could help them practice applying and combining different topics to solve more advanced problems. Additionally, instructors could ensure that exercises cover all the topics assessed in the final exam and conduct item analysis to improve the quality of the exam so it is inclusive and equitable for all students. These types of actionable insights can benefit educators as they are increasingly confronted with complex challenges and motivate them to practice data-informed decision-making. As a result, student learning may be improved. So far, prediction models were indeed used for producing such actionable insights for either instructors or students, and XAI was used along with the final product (e.g., Afzaal et al., 2021; Er et al., 2020; Jang, Choi, & Kim, 2022). We went an important step forward and suggested to

involve stakeholders as an integral part of the design of the prediction models. Engaging stakeholders in the development process of prediction models may improve the models by adding their knowledge and expertise (Hershkovitz & Ambrose, 2022). It can also reciprocally promote the engaged stakeholders, as they would better understand how prediction models work, and how such models could improve their teaching, which in turn can enhance their trust in the models. We did so by including the instructor as an integral, equal partner in the development of the prediction model, hence taking a co-design approach to prediction model construction. This approach is in line with the emerging practice of participatory design in learning analytics (often referred to as human-centered learning analytics) (Dollinger et al., 2019; Prieto-Alvarez, Martinez-Maldonado, & Dirndorfer Anderson, 2018). A recent review of such approaches (Sarmiento & Wise, 2022) revealed that co-design was carried out before or after the learning analytics design in most cases. It also showed co-design was mostly implemented as a means for generation ideas or evaluating early ideas even when it was done during the design process. The review found only a few cases in which co-design was implemented as a co-development of learning analytics, but those cases didn't employ co-design at the model construction level. Therefore, our co-design approach is novel and unique, and we encourage other learning analytics endeavors to implement it.

Our study demonstrated how to open the black-box predictive models and invite the end-users of those models to share their judgments of the prediction. By involving those who are experts both in the subject matter and in teaching, such as instructors and learning scientists, in evaluating the prediction, we overcame the limitations of the conventional evaluation metrics. Traditionally, the regression models are evaluated by merely aggregative accuracy metrics, like MAE or RMSE. However, as demonstrated in our study, such metrics don't consider the course context under

which learning occurs and therefore are not sufficient in determining the validity of predictive models in educational settings. It is essential to consider educators' judgment of the models' outputs. By leveraging educators' judgment and expertise, we can not only enhance the performance of the predictive models but also improve their trustworthiness to the end users, which may optimize the teaching effort and improve the learning outcome.

We believe that besides these lessons for researchers, our findings also have some meaningful implications for both instructors and students. First, instructors should pay closer attention to the heterogeneity in their classrooms, and to the fact that seemingly obvious measures—like course exercise scores—are not necessarily fully aligned with student learning and understanding. High-performing and low-performing students may respond very differently towards course exercise and still demonstrate similar (low) performance; therefore, whether an exercise is "easy" or "difficult" is not to be determined merely by its scores, and instructors should be aware to it. A possible solution to this is to provide students with a host of exercises and to allow them some degree of choice, in a way that will enable every student to practice successfully and to thrive. For students, our findings suggest the need to reflect on their learning process throughout it, and to pay close attention to the way they respond to various challenges in which they tackle; these challenges may be related to, e.g., content, skills, problem presentation, or self-motivation, and they should be encouraged to look up close on their learning and to make such distinctions. Following this recognition, they should be encouraged to seek help addressing their main challenges, and therefore to better their learning, and to improve themselves as learners. At large, these modifications of teaching and learning will help in improving education.

6 LIMITATIONS AND FUTURE WORK

Our study provides an initial step to explore the trustworthiness and transparency of AI in educational settings based on XAI in combination with instructors' contextual knowledge for modeling student learning prediction processes. We employed the Kernel SHAP method to explain how the model made the prediction at both the global and local levels. The SHAP value calculation can become computationally intractable for larger feature sets given the number of possible coalitions increases exponentially with the number of features. Its time complexity becomes $O(kM2^M)$ when taking the average of k samples of M features. Despite the limitations of the relatively small sample size, this study presents a holistic analytical process in terms of explaining the underlying reasoning of AI models' predictions and incorporating instructors' contextual knowledge to improve the models' performance and trustworthiness. This whole process has significant practical implications in terms of guiding instructors to investigate and better apply the insights generated from AI models in their instructional processes and researchers in the corresponding field to conduct complementary research to generalize the findings in our study. Furthermore, with a specific focus on AI trustworthiness and transparency, our research can also have the potential to improve confidence in employing AI technologies in some high-stakes areas other than education, such as healthcare and biomedical engineering domains.

To better realize the potential of AI in education, it is important to evaluate the AI models in actual working environments constrained by various protocols, which need a deeper level of human-machine interaction and collaboration. As such, the future research direction lies in the following two directions: (1) exploring technologies to support more interactive human and AI interactions for achieving trustworthy and transparent AI in educational settings; (2) exploring

explainable AI technologies that are capable of dynamically incorporating subject-matter experts' contextual insights during the learning process for generating more accountable interventions aimed to help students accomplish their learning goals. It is hoped that the research in these aspects will further promote the applications of AI in education to enhance student learning experience and improve their learning outcomes.

Statements and Declarations

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

Afzaal, M., Nouri, J., Zia, A., Papapetrou, P., Fors, U., Wu, Y., ... & Weegar, R. (2021, June). Generation of automatic data-driven feedback to students using Explainable Machine Learning. In *International Conference on Artificial Intelligence in Education* (pp. 37-42).

Al-Shabandar, R., Hussain, A. J., Liatsis, P., & Keight, R. (2019). Detecting at-risk students with early interventions using machine learning techniques. *IEEE Access*, 7, 149464-149478.

Anwar, M. (2021). Supporting privacy, trust, and personalization in online learning. *International Journal of Artificial Intelligence in Education, 31*(4), 769-783.

Azeta, A. A., Ayo, C. K., Atayero, A. A., & Ikhu-Omoregbe, N. A. (2009, January). A case-based reasoning approach for speech-enabled e-learning system. In *2009 2nd*

International Conference on Adaptive Science & Technology (ICAST) (pp. 211-217). IEEE.

Ben-Zadok, G., Leiba, M., & Nachmias, R. (2010). Comparison of online learning behaviors in school vs. at home in terms of age and gender based on log file analysis. *Interdisciplinary Journal of E-Learning and Learning Objects*, 6(1), 305-322.

Botelho, A. F., Baker, R. S., & Heffernan, N. T. (2019, July). Machine-learned or expert-engineered features? Exploring feature engineering methods in detectors of student behavior and affect. In *The twelfth international conference on educational data mining.*

Cheema, J. R., & Sheridan, K. (2015). Time spent on homework, mathematics anxiety and mathematics achievement: Evidence from a US sample. *Issues in Educational Research, 25*(3), 246-259.

Ciolacu, M. I., & Svasta, P. (2021, April). Education 4.0: AI empowers smart blended learning process with Biofeedback. In *2021 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1443-1448). IEEE.

Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. *arXiv preprint arXiv:1807.00154*.

De Jong, R., Westerhof, K. J., & Creemers, B. P. (2000). Homework and student math achievement in junior high schools. *Educational research and Evaluation, 6*(2), 130-157.

Dollinger, M., Liu, D., Arthars, N., & Lodge, J. M. (2019). Working together in learning analytics towards the co-creation of value. *Journal of Learning Analytics*, 6(2), 10-26.

Duan, X., Wang, C., & Rouamba, G. (2022). Designing a Learning Analytics Dashboard to Provide Students with Actionable Feedback and Evaluating Its Impacts. In *CSEDU (2)* (pp. 117-127).

Er, E., Gomez-Sanchez, E., Bote-Lorenzo, M. L., Dimitriadis, Y., & Asensio-Pérez, J. I. (2020). Generating actionable predictions regarding MOOC learners' engagement in peer reviews. *Behaviour & Information Technology*, *39*(12), 1356-1373.

Fan, H., Xu, J., Cai, Z., He, J., & Fan, X. (2017). Homework and students' achievement in math and science: A 30-year meta-analysis, 1986–2015. *Educational Research Review*, *20*, 35-54.

Fernández-Alonso, R., Suárez-Álvarez, J., & Muñiz, J. (2015). Adolescents' homework performance in mathematics and science: personal factors and teaching practices. *Journal of educational psychology*, *107*(4), 1075.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines*, *28*(4), 689-707.

Goel, Y., & Goyal, R. (2020). On the effectiveness of self-training in mooc dropout prediction. *Open Computer Science*, *10*(1), 246-258.

Hänsch, N., Schankin, A., Protsenko, M., Freiling, F., & Benenson, Z. (2018). Programming experience might not help in comprehending obfuscated source code efficiently. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)* (pp. 341-356).

Heras, S., Palanca, J., Rodriguez, P., Duque-Méndez, N., & Julian, V. (2020). Recommending learning objects with arguments and explanations. *Applied Sciences*, *10*(10), 3341.

Hershkovitz, A., & Nachmias, R. (2009). Consistency of Students' Pace in Online Learning. *International Working Group on Educational Data Mining*.

Hershkovitz, A., & Ambrose, A. (2022). Insights of Instructors and Advisors into an Early Prediction Model for Non-Thriving Students. *Journal of Learning Analytics,* 9(2), 202-217.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021, March). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 624-635).

Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and eXplainable AI. *Education and Information Technologies,* 1-35.

Jiang, Y., Bosch, N., Baker, R. S., Paquette, L., Ocumpaugh, J., Andres, J. M., ... & Biswas, G. (2018, June). Expert feature-engineering vs. deep neural networks: which is better for sensor-free affect detection?. In *International conference on artificial intelligence in education* (pp. 198-211). Springer, Cham.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3, 100074.

Kim, W. H., & Kim, J. H. (2020). Individualized AI tutor based on developmental learning networks. *IEEE Access*, *8*, 27927-27937.

Kitsantas, A., Cheema, J., & Ware, H. W. (2011). Mathematics achievement: The role of homework and self-efficacy beliefs. *Journal of Advanced Academics*, *22*(2), 310-339.

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014, October). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 workshop on analysis of large-scale social interaction in MOOCs* (pp. 60-65).

Knowles, B., & Richards, J. T. (2021, March). The sanction of authority: Promoting public trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 262-271).

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121-204.

Lee, U. J., Sbeglia, G. C., Ha, M., Finch, S. J., & Nehm, R. H. (2015). Clicker score trajectories and concept inventory scores as predictors for early warning systems for large STEM classes. *Journal of Science Education and Technology*, *24*(6), 848-860.

Levin, N. A. (2021). Process Mining Combined with Expert Feature Engineering to Predict Efficient Use of Time on High-Stakes Assessments. *Journal of Educational Data Mining,* 13(2), 1-15.

Li, J., Li, H., Majumdar, R., Yang, Y., & Ogata, H. (2022, March). Self-directed Extensive Reading Supported with GOAL System: Mining Sequential Patterns of Learning Behavior and Predicting Academic Performance. In LAK22: 12th International Learning Analytics and Knowledge Conference (pp. 472-477).

Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers, Environment and Urban Systems*, *96*, 101845.

Lu, S., Chen, R., Wei, W., Belovsky, M., & Lu, X. (2021). Understanding Heart Failure Patients EHR Clinical Features via SHAP Interpretation of Tree-Based Machine Learning Model Predictions. In *AMIA Annual Symposium Proceedings* (Vol. 2021, p. 813). American Medical Informatics Association.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & education*, *54*(2), 588-599.

Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, *2021*.

Marras, M., Vignoud, J. T. T., & Kaser, T. (2021). 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).

Matcha, W., Gašević, D., & Pardo, A. (2019). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies*, *13*(2), 226-245.

Mokhtari, K. E., Higdon, B. P., & Başar, A. (2019, November). Interpreting financial time series with SHAP values. In *Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering* (pp. 166-172).

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071-22080.

Nazaretsky, T., Cukurova, M., & Alexandron, G. (2022, March). An Instrument for Measuring Teachers' Trust in AI-Based Educational Technology. In *LAK22: 12th international learning analytics and knowledge conference (pp. 56-66)*.

Ndiyae, N. M., Chaabi, Y., Lekdioui, K., & Lishou, C. (2019, March). Recommending system for digital educational resources based on learning analysis. In *Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society* (pp. 1-6).

Prieto-Alvarez, C. G., Martinez-Maldonado, R., & Anderson, T. D. (2018). Co-designing learning analytics tools with learners. In *Learning Analytics in the Classroom* (pp. 93-110)

Pejić, A., & Molcer, P. S. (2021). Predictive machine learning approach for complex problem solving process data mining. *Acta Polytechnica Hungarica,* 18(1), 45-63.

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458-472.

Rong, Q., Lian, Q., & Tang, T. (2022). Research on the Influence of AI and VR Technology for Students' Concentration and Creativity. *Frontiers in Psychology*, *13*.

Rotelli, D., & Monreale, A. (2022, March). Time-on-Task Estimation by data-driven Outlier Detection based on Learning Activities. In *LAK22: 12th International Learning Analytics and Knowledge Conference (pp. 336-346)*.

Sarmiento, J. P., & Wise, A. F. (2022, March). Participatory and Co-Design of Learning Analytics: An Initial Review of the Literature. In *LAK22: 12th International Learning Analytics and Knowledge Conference (pp. 535-541).*

Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4), 1-31.

Swamy, V., Radmehr, B., Krco, N., Marras, M., & Käser, T. (2022). Evaluating the Explainers: Black-Box Explainable Machine Learning for Student Success Prediction in MOOCs. *arXiv preprint arXiv:2207.00551*.

Syed, M., Anggara, T., Lanski, A., Duan, X., Ambrose, G. A., & Chawla, N. V. (2019, March). Integrated closed-loop learning analytics scheme in a first year experience course. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 521-530).

Szafir, D., & Mutlu, B. (2013, April). ARTFul: adaptive review technology for flipped learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1001-1010).

Takami, K., Dai, Y., Flanagan, B., & Ogata, H. (2022, March). Educational Explainable Recommender Usage and its Effectiveness in High School Summer Vacation

Assignment. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 458-464).

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, *31*(2), 447-464.

Thornton, L., Knowles, B., & Blair, G. (2021, March). Fifty Shades of Grey: In Praise of a Nuanced Approach Towards Trustworthy Design. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 64-76).

Trautwein, U. (2007). The homework–achievement relation reconsidered: Differentiating homework time, homework frequency, and homework effort. *Learning and instruction*, *17*(3), 372-388.

Vincent-Lancrin, S., & van der Vlies, R. (2020). Trustworthy artificial intelligence (AI) in education: Promises and challenges.

Zhang, J. H., Zou, L. C., Miao, J. J., Zhang, Y. X., Hwang, G. J., & Zhu, Y. (2020). An individualized intervention approach to improving university students' learning performance and interactive behaviors in a blended learning environment. *Interactive Learning Environments*, *28*(2), 231-245.

Zhang, M., Guo, H., & Liu, X. (2021). Using Keystroke Analytics to Understand Cognitive Processes during Writing. *International Educational Data Mining Society.*