Expert Systems

WILEY

Expert Systems

ORIGINAL ARTICLE OPEN ACCESS

Utilising Explainable AI to Enhance Real-Time Student Performance Prediction in Educational Serious Games

Manuel J. Gomez | Álvaro Armada Sánchez | Mariano Albaladejo-González | Félix J. García Clemente | José A. Ruipérez-Valiente

Department of Information and Communications Engineering, University of Murcia, Murcia, Spain

Correspondence: Mariano Albaladejo-González (mariano.albaladejog@um.es)

Received: 20 March 2024 | **Revised:** 24 October 2024 | **Accepted:** 22 January 2025

Funding: This work was supported by the Fundación Séneca, Grant/Award Number: 21795/FPI/22, 21948/JLI/22 and 22238/PDC/23; Instituto Nacional de Ciberseguridad, Grant/Award Number: CDL-TALENTUM.

Keywords: artificial intelligence | just-in-time interventions | learning analytics | machine learning | xAI

ABSTRACT

In recent years, serious games (SGs) have emerged as a powerful tool in education by combining pedagogy and entertainment, facilitating the acquisition of knowledge and skills in engaging environments. SGs enable the collection of valuable interaction data from students, allowing for the analysis of student performance, with artificial intelligence (AI) playing a key role in processing this data to make informed inferences about their knowledge and skills. However, the lack of explainability in AI models represents a significant challenge. This research aims to develop an interpretable model for predicting students' performance in real-time while playing an SG by: (1) calculating the performance of an interpretable prediction model of task completion in an SG and (2) demonstrating the application of the interpretable model for just-in-time (JIT) classroom interventions. Our results show that we are able to predict students' task completion in real-time with a balanced accuracy result of 77.21% after a short playtime has elapsed. In addition, an explainable artificial intelligence (XAI) approach has been applied to ensure the interpretability of the developed models. This approach supports personalised learning experiences, unlocks AI benefits for non-technical users, and maintains transparency in education.

1 | Introduction

In recent years, the use of games in educational contexts has significantly increased. Digital games provide an additional way for students to develop cognitive, spatial, and motor skills; help improve information and communication technology knowledge; teach complex problem-solving; and increase creativity, all while addressing topics that might be perceived as too complicated in a traditional classroom setting (Papanastasiou et al. 2017). Specifically, the potential of serious games (SGs)—games that do not have entertainment, enjoyment, or fun as their main purpose (Laamarti, Eid, and El Saddik 2014)—is particularly relevant, as they offer a unique blend of entertainment and pedagogy. SGs can provide an excellent context that not only facilitates the acquisition and assessment of knowledge and skills but also allows for a detailed examination of environments free from the usual constraints of time and space (Bellotti, Berta, and De Gloria 2010). In the field of education, there is significant enthusiasm surrounding game-based assessment (GBA) due to the apparent limitations of conventional assessment methods in capturing students' knowledge, skills, and attributes (de Klerk and Kato 2017).

One of the key benefits of utilising SGs in educational settings is the wealth of data that can be collected from these interactive experiences, providing a great opportunity to make inferences

and assessments in ways that are not possible in traditional testing (Gomez, Ruipérez-Valiente, and Clemente 2023). The scope of the collected data can range from measuring individual skills at a granular level to evaluating attitudes at a larger scale. Moreover, the collected data can be intentionally designed to assess various aspects, such as knowledge, attitudes, skills, or behaviour (Smith, Blackmore, and Nesbitt 2015). Data from SGs can also be used by teachers to provide targeted assistance and support to students precisely when they need it. Specifically, just-in-time (JIT) instruction occurs when information, skill demonstration, or other necessary instruction is delivered on the spot at the time it is required, ensuring that information is available for immediate application in the relevant context (Anderson and Wood 2009).

The use of SGs in education does not necessarily imply teacher disengagement from teaching. Like any other learning situation, students require the guidance of their teachers during gameplay. It is the teacher's responsibility to ensure that all students are progressing through the game and successfully achieving both the game goals and the learning objectives (Bado 2022). However, monitoring students in real-time poses challenges that encompass several key aspects educators face in educational environments, such as conducting meaningful data analysis and identifying when and how to intervene based on the collected data. The use of artificial intelligence (AI) can support teachers in this task. In particular, monitoring students' performance in SGs generates a substantial amount of information that needs to be processed, and Machine Learning (ML) models can handle this data and identify patterns that allow us to recognise behaviours (Marín-Morales et al. 2021). That being said, the majority of existing research focuses on post hoc analyses or predicting overall student performance after finishing the game. Few studies have attempted to provide real-time insights that could inform JIT interventions.

Although ML models have demonstrated strong predictive capabilities in educational contexts, explainability remains an inherent problem of the latest techniques (e.g., ensembles or Deep Neural Networks) (Arrieta et al. 2020), and the increasing utilisation of ML models has led to a growing demand for transparency in AI (Preece et al. 2018). In this sense, explainable AI (XAI) proposes creating a suite of techniques that produce more explainable models while maintaining high performance. This enables end users to understand, appropriately trust, and effectively manage the emerging generation of AI models (Gunning 2017). In this way, XAI tools empower non-technical users (such as teachers) to interpret AI-generated insights and recommendations, enabling them to make informed decisions. Yet, there is a clear gap in research focusing on the combination of real-time predictive models and XAI in SGs. To the best of our knowledge, no previous studies have explored how interpretable models can be used to deliver JIT interventions during gameplay by predicting students' performance.

In this research, we aim to address these gaps by building an explainable model capable of predicting students' level outcomes in real-time while playing an SG. With this purpose in mind, we design a set of features derived from students' interaction with *Shadowspect*, a game designed as a formative assessment tool to assess mathematical content standards. Using these features, we build a set of ML models that aim to predict the students' performance in real-time while ensuring that the models remain explainable. Unlike previous studies that have mainly focused on post-game performance predictions, our work makes early predictions during gameplay. In addition, we make results explainable and immediately applicable to JIT interventions. This real-time, interpretable feedback is a key contribution, as it enables educators to intervene promptly based on students' progress. Specifically, we have the following objectives:

1. **Configure and evaluate interpretable models of task completion at different time windows in an SG**. We aim to assess the model's performance by measuring its predictive accuracy and evaluating its ability to anticipate students' progress within the game environment. Furthermore, we will address the interpretability of the AI models to ensure that end users can interpret the results.
2. **Demonstrate the application of the interpretable model for JIT classroom interventions**. Our final objective is to showcase the practical application of the interpretable prediction model for real-time classroom interventions. We will conduct a use case that illustrates how our model can be applied in a real scenario.

The rest of the paper is structured as follows: Section 2 reviews background literature on SGs and assessment, AI models and XAI. Section 3 describes the methodology followed to conduct the research, as well as the game and the data collection used. Next, Section 4 presents the results, including the models developed, their interpretability analysis, and finally the case study conducted. Then, we finalise the paper with a discussion in Section 5 and conclusions and future work in Section 6.

2 | Related Work

The concept of games dates back to ancient civilisations and is recognised as a fundamental aspect of human societies throughout history (Laamarti, Eid, and El Saddik 2014). In particular, the versatility and adaptability of SGs make them a valuable tool across various contexts and domains. First of all, SGs have gained significant popularity in educational settings. For example, Ruiperez-Valiente et al. (2020) used "The Radix Endeavor" (an inquiry-based online game for STEM learning) in K-12 classrooms as part of a pilot study conducted in numerous schools. Additionally, SGs have also been proposed as a potential method for employee selection by improving the user experience, and the use of games in the workplace is a growing phenomenon, with SGs being increasingly used as evaluative tools (Al Qallawi and Raghavan 2022). Larson (2020) conducted a literature review on SGs and gamification in corporate environments, finding that the use of SGs is becoming increasingly prevalent. Moreover, SGs have been considered positive and innovative solutions for addressing contemporary issues in organisations, including meeting the needs of modern learners within the corporate context. In healthcare, SGs, particularly adventure and shooter games, play an important role in education, prevention and rehabilitation (Wiemeyer and Kliem 2012). In this regard, one of the most challenging areas is modelling simulations for

medical training. 'CancerSpace', developed by the National Cancer Institute and Oak Ridge Institute for Science and Education, is an SG that aims to facilitate cancer screening and consequently increases cancer-screening rates in federally qualified health centres (Swarz et al. 2010).

Moreover, predicting students' performance in educational environments has gained considerable attention due to its potential to transform educational practices and enhance learning outcomes. Rastrollo-Guerrero, Gómez-Pulido, and Durán-Domínguez (2020) conducted a literature review of 70 papers to examine techniques and objectives for predicting students' performance, noting a strong tendency to focus on university-level predictions (around 70% of the analysed articles). However, the authors also highlighted the need to apply these predictions at the school level, which would help identify low-performing students at earlier ages. Regarding SGs, there have been several studies attempting to analyse students' data to predict their performance. For example, research presented by Illanas Vila et al. (2013) aimed to predict students' performance in translating foreign languages by collecting data from 55 students in an SG and building neural network models. Although the results obtained were positive, the authors acknowledged certain limitations, such as the potential bias in their data set. Additionally, Kickmeier-Rust (2018) conducted a simulated study based on existing datasets using a multidimensional domain and learner models to add information about the nature of a learning domain. Furthermore, Abeyrathna et al. (2019) built a multi-label classifier using in-game data and player information to predict student proficiency in a quantum cryptography SG. More recent work by Hooshyar et al. (2023) attempted to predict early student performance using only 50% of learners' action sequences, achieving a relative error of less than 8%. Finally, other studies have used ML models to analyse students' interactions and predict their performance using in-game data (Yuhana et al. 2017; Loh, Sheng, and Li 2015; Lee et al. 2023; Alonso-Fernández et al. 2020). However, while these studies focused on performance prediction, none have demonstrated how to make their results explainable or immediately applicable to JIT interventions.

There are many institutions already using AI technologies to shape and plan the delivery of education (Zawacki-Richter et al. 2019). The use of AI has become a focal point for innovation and competitive advantage, with applications anticipated in areas such as learner profiling, intelligent tutoring systems, assessment, and personalised learning (Farrow 2023). XAI emerges to address the 'black box' perception non-technical users often have about AI, which can seem 'humanly inexplicable'. As AI becomes more prevalent in education, XAI should help educators and learners understand the algorithms that influence the learning process. Previous research has examined the use of XAI in educational contexts. For example, Tao et al. (2020) presented an explainable multi-view game cheating detection framework driven by XAI.

Regarding the use of XAI for predicting student performance, Chitti, Chitti, and Jayabalan (2020) noted that the prediction models generated are often complex and not interpretable, making it difficult to understand why and how predictions are made based on the results. Thus, model interpretability has become increasingly important. Alamri and Alharbi (2021) conducted a literature review investigating explainable models for predicting student performance from 2015 to 2020. Their findings revealed that the predictors used to train these models primarily consisted of a combination of socio-economic features and pre-course performance features. However, the review also highlighted that the potential of utilising e-learning analytics data as a source for explainable student performance models has not been fully explored. Nevertheless, we found some studies that used XAI to create interpretable student performance models. For example, Jang et al. (2022) applied several ML models to predict performance and verify whether at-risk students could be identified using selected features, providing helpful information to each student through XAI techniques. Regarding SGs, Berger and Müller (2021) designed a rule-based, short-term decision-making algorithm that reports game progress, demonstrating its suitability for creating adaptive SGs.

Table 1 provides a detailed comparison of previous studies on student performance prediction and XAI applications in SGs. Our research stands out from existing literature on predicting student performance in SGs by addressing two crucial aspects: real-time prediction and model explainability. Although Berger and Müller (2021) also tackled model interpretability and real-time prediction, their approach was entirely different, as they developed a rule-based model to predict in-game progress for adaptive SGs. In contrast, our work focuses on predicting student performance using ML models, with an emphasis on interpreting those predictions in real-time.

As observed, previous studies have attempted to forecast student performance in games, but these predictions have typically been made after the completion of the activity or game. Our research aims to push the boundaries by making real-time predictions using in-game data. By incorporating XAI techniques, we aim to shed light on the underlying factors and decision-making processes influencing student performance predictions.

3 | Methodology

Next, we present the SG *Shadowspect* along with our data collection in Section 3.1, followed by a detailed description of the complete process used to conduct the study in Section 3.2.

3.1 | Context and Dataset

In our research, we utilised *Shadowspect*, a 3D geometry game specifically designed to evaluate math core standards, including the visualisation of relationships between 2D and 3D objects. This allows teachers to integrate it into their core math curriculum. The game enables students to create composite figures using primitive shapes (e.g., pyramids, cones) and silhouettes from different perspectives. Players can manipulate shapes, change perspectives, and receive feedback on how well they match the silhouettes. An example of a puzzle being solved is shown in Figure 1. The current version of the game consists of 30 levels, divided into nine tutorial, nine intermediate and 12 advanced levels. While the tutorial levels focus on teaching

TABLE 1 | Detailed comparison of existing studies on student performance prediction and XAI applications in SGs.

Study	Prediction goal	Methods	Data source	Dataset users	Interpretable model	Real-time prediction
Illanas Vila et al. (2013)	Predicting student performance in foreign languages	Neural networks	In-game data	55	✗	✗
Kickmeier-Rust (2018)	Predicting student performance in game-based scenarios	Linear regression model	Math competencies	912	✓	✗
Abeyrathna et al. (2019)	Predicting student proficiency in quantum cryptography	Support vector machine	In-game and player data	150	✗	✗
Loh, Sheng, and Li (2015)	Predicting expert-novice performance differences	Partial least squares discriminant analysis	In-game data	62	✗	✗
Lee et al. (2023)	Predicting student posttest math knowledge scores	Seven ML models	In-game data	359	✗	✗
Yuhana et al. (2017)	Predicting student performance in math skills	Five ML models, one rule-based classifier	In-game answers and player data	160	✗	✗
Hooshyar et al. (2023)	Predicting student performance at early stages	Eight ML models, one deep learning (DL) model	In-game features	427	✗	✓
Alonso-Fernández et al. (2020)	Predicting student knowledge given as post-test results	Decision trees, naïve bayes, logistic regression	In-game data	227	✓	✗
Berger and Müller (2021)	Predicting in-game progress for adaptive SGs	Rule-based algorithm	In-game data	80	✓	✓
Our work	Predicting student-level performance in real-time	Seven ML models	In-game data	322	✓	✓

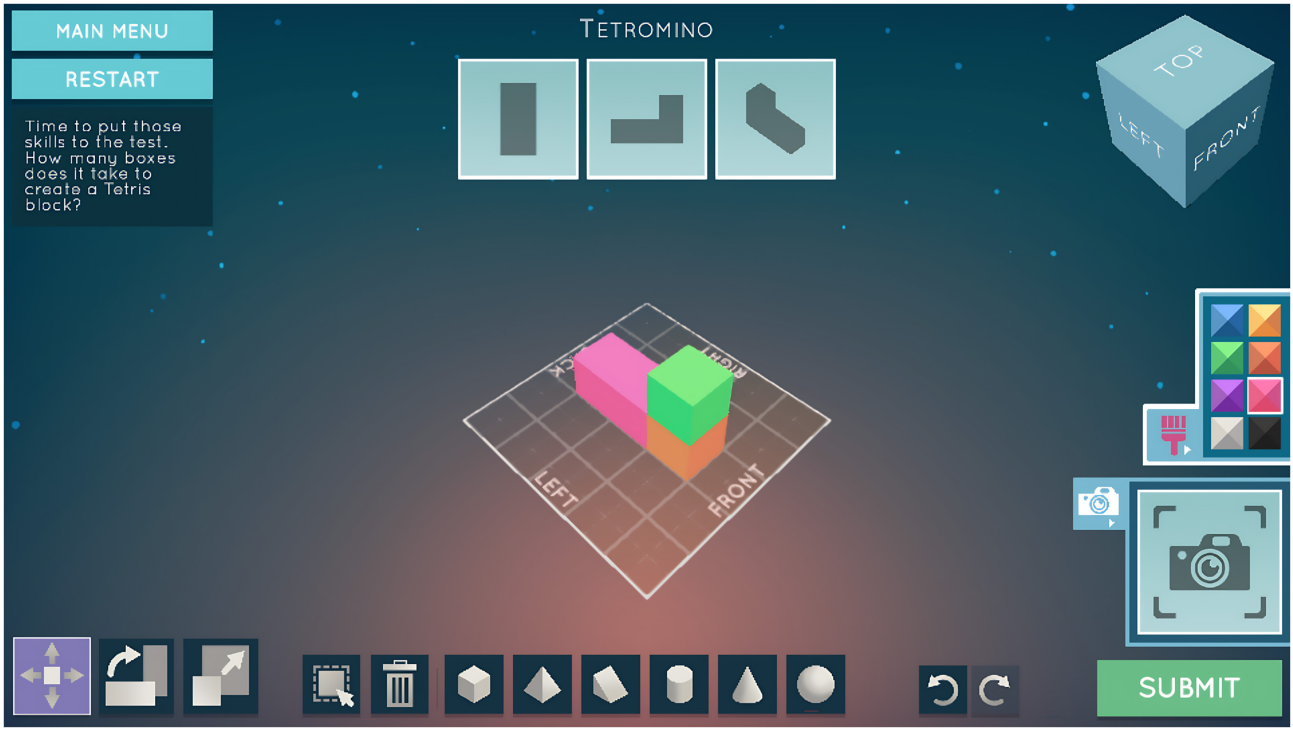

FIGURE 1 | Puzzle example in *Shadowspect*.

basic functionality, the intermediate and advanced levels offer more independence and challenging puzzles for experienced players.

For the data collection process, the team recruited seven teachers who used the games with students from seventh to tenth grade. The final dataset used for this research consists of approximately 428,000 events performed by a total of 322 students (with an average of 1320 events per student). These events were recorded over a period of 260 h, equivalent to an average of 0.82 h per student.

3.2 | Training and Explicability Procedure

We can divide the methodology into two main blocks: performance prediction model development and model interpretability and explanations. The first block describes the methodology used to build the ML models for performance prediction and consists of two stages: feature engineering and model training. In the feature engineering stage, a set of features was developed to assist in prediction, while the model training stage details how the ML models were trained. The second block focuses on the methodology used to enhance model interpretability, with the approach adapted according to the model's inherent interpretability. We can see a diagram illustrating the complete methodology in Figure 2.

3.2.1 | Feature Engineering

The first step was to design the features used for predicting users' performance in *Shadowspect*, specifically whether a user would successfully complete the level being played. To achieve this, we employed a multi-prediction approach based on distinct time intervals derived from the average time required to complete each in-game puzzle. These intervals were set at 25%, 50%, and 75% of the average completion time, allowing us to monitor player progress and provide timely feedback or interventions as needed.

We designed a set of features crucial for predicting user success, categorised into three groups: **user features**, **puzzle features**, and **attempt features**. **User features** provide insights into the user's overall performance and interaction patterns, helping us understand their ability and strategies in the game. **Puzzle features** cover a set of data from the different levels within the game, allowing us to capture the unique characteristics and difficulty of each level, which is essential for accurate predictions. Finally, **attempt features** provide specific information about what the user is doing in that specific attempt, giving us a detailed view of the actions and decisions during gameplay. Using this comprehensive feature design allows us to consider both user general behaviour and the specifics of each gaming scenario. Table 2 presents the user features, Table 3 displays the puzzle features, and finally, Table 4 shows the attempt features. These features aim to summarise all aspects of users' interactions with *Shadowspect* during puzzle-solving attempts. It is important to note that **puzzle features** and **user features** do not vary with different time intervals since they aim to provide an overview of the user and the puzzle being played. Therefore, the features that vary with time intervals will be **attempt features**, as they provide information about the current game session.

Regarding the `user_elo` and `puzzle_elo` features, they are both calculated using an adapted version of the original Elo algorithm, which was initially designed as a method to rank chess players (Elo 2008). In our context, we consider the student and the puzzle as the opponents in our game, and each student's

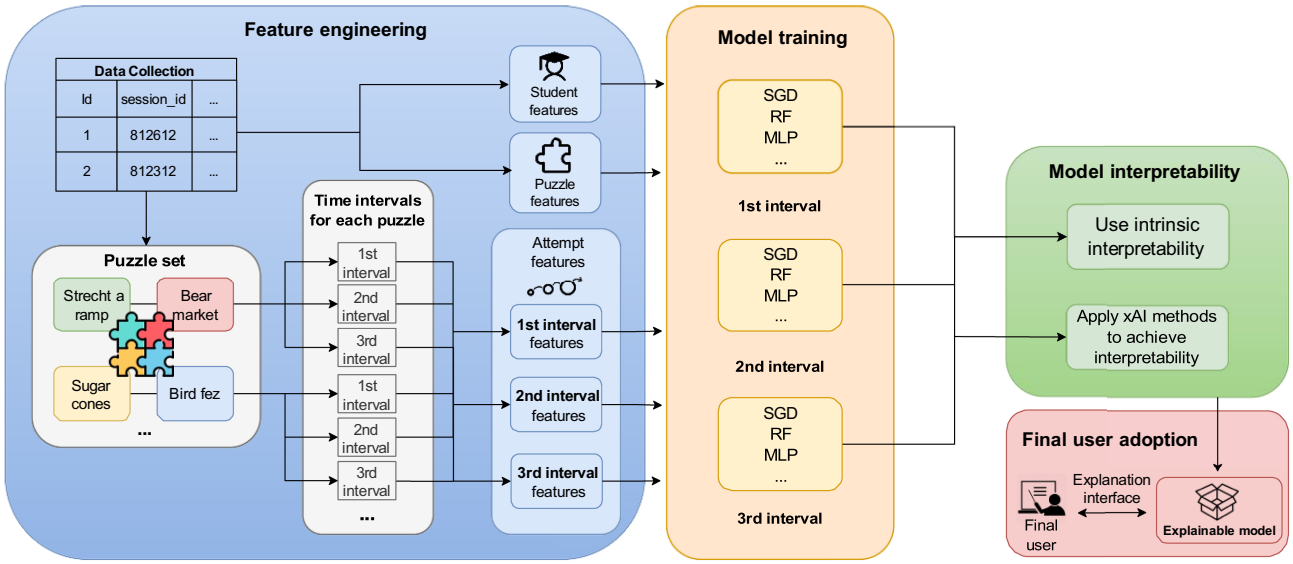

FIGURE 2 | Methodology diagram.

TABLE 2 | User features.

Feature	Description
`percentage_tutorial`	Percentage of completed puzzles in the tutorial category.
`percentage_intermediate`	Percentage of completed puzzles in the intermediate category.
`percentage_advanced`	Percentage of completed puzzles in the advanced category.
`attempts_per_puzzle`	The average number of solution attempts needed by the user to solve a puzzle.
`user_elo`	Elo ranking obtained by facing each user against different puzzles.

TABLE 3 | Puzzle features.

Feature	Description
`puzzle_difficulty`	Calculated using the average time it takes to complete the puzzle, the average number of actions to solve it, the percentage of abandonments, and the percentage of incorrect checks when attempting to solve it are normalised separately with respect to the same metrics obtained from the rest of the puzzles. Once normalised, they are aggregated and normalised again with respect to the distribution of puzzles.
`puzzle_elo`	Elo ranking obtained by facing each puzzle against different users.
`puzzle`	The string containing the name of the puzzle.

attempt at solving a puzzle is considered a match. A comprehensive description of how the adapted algorithm works can be checked in previous work (Ruipérez-Valient et al. 2022).

3.2.2 | Model Training

For training our models, we have considered the following algorithms:

- **Adaboost**: An ensemble method that combines weak learners to create a strong learner by iteratively adjusting weights.
- **Decision Tree (DT)**: Creates a tree-like model of decisions by recursively splitting data based on different feature conditions.
- **K-Nearest Neighbours (KNN)**: Assigns a label to a new data point based on the labels of its k nearest neighbours.
- **Multi-Layer Perceptron (MLP)**: A type of feedforward neural network with multiple layers to model complex non-linear relationships in data.
- **Random Forest (RF)**: Ensemble method combining multiple decision trees to provide reliable predictions.
- **Stochastic Gradient Descent (SGD)**: Efficient optimisation algorithm using small random subsets of training data, suitable for large datasets and online learning.

TABLE 4 | Attempt features.

Feature	Description
n_events	Total number of events generated by the user.
n_breaks	Number of user's idle moments (15 s without generating any events).
n_snapshot	Number of screenshots taken by the user.
n_rotate_view	Number of camera rotations performed by the user.
n_manipulation_events	Number of manipulation events generated by the user.
n_check_solution	Number of attempts to check the solution.
best_submit	Best puzzle submission rate obtained, determined by dividing the number of matched silhouettes by the total number of correct silhouettes.

- **Support Vector Classifier (SVC)**: Constructs hyperplanes to separate data into different classes, effective in high-dimensional spaces and for complex decision boundaries.

Regarding the data preprocessing, it is worth mentioning the separation of the dataset into training and testing sets. In this work, we needed to consider an additional condition beyond simply separating the data. We wanted to ensure that attempts from the same user fell into the same dataset to prevent the model from learning based on a user's specific behaviour and then predicting other attempts from the same user. This guarantees that the model is capable of generalising well to user behaviour across different individuals. Therefore, we decided to randomly select 70% of the users as training users and the remaining 30% as testing users. The attempts corresponding to the training users were included in the training dataset, while the attempts made by the testing users were included in the testing dataset. In addition, we applied one-hot encoding to the `puzzle` feature because some of the ML models used are unable to handle categorical data.

For model training and configuration, we employed ten-fold cross-validation and used balanced accuracy as the performance metric. This metric calculates the percentage of correctly classified positive and negative instances and then averages these percentages. It is highly useful in scenarios with imbalanced data, as it assigns equal importance to the accuracy of both the majority and minority classes. In our case, we are dealing with imbalanced data, as the number of instances corresponding to successes is nearly twice that of failures; thus, we prioritise balanced accuracy as our primary evaluation metric. In addition to the balanced accuracy, we also reported the F1 score, Matthews correlation coefficient (MCC), precision, sensitivity and specificity. After training each model configuration separately, we selected the models and configurations associated with the algorithms that achieved the best average balanced accuracy for each time interval.

3.2.3 | Model Interpretability and Explanations

Once the best model was selected, we aimed to enhance its interpretability, ensuring that the model's predictions were fully explainable. If the chosen model was inherently interpretable, we would utilise that interpretability to explain the model's predictions. For example, if the best model happened to be a *decision tree*, we could easily interpret its decision-making process by examining the sequence of split rules and feature importance. On the other hand, if the chosen model was not interpretable, we sought to apply XAI methods to achieve interpretability.

XAI methods can be categorised based on different criteria. The first criterion is 'intrinsic vs. post hoc', where interpretability is achieved either by constraining the complexity of the ML model (intrinsic) or by using methods that analyse the model after training (post hoc). The second criterion is 'model-specific vs. model-agnostic'. Model-specific interpretation tools are limited to certain model classes, like interpreting regression weights in linear models. Model-agnostic tools, on the other hand, can be used with any ML model and are applied after the model has been trained. Lastly, the third criterion is 'local vs. global' interpretation. Local methods explain individual predictions, while global methods provide explanations for the overall behaviour of the entire model.

In this particular scenario, our intention was to use a post hoc, model-agnostic, and local interpretability method. We specifically opted for a post hoc approach because the chosen model was not inherently explainable. Furthermore, we aimed for it to be model-agnostic, as using a method tied to a particular model might not be compatible with the best-performing model. Lastly, we chose a local method, as our primary focus was on explaining individual predictions. The selected method for adding interpretability to a 'black box' model was *SHAP*. The primary reason for choosing *SHAP* was its capability to provide local interpretability. Additionally, *SHAP* values can be aggregated globally, offering insights into the model's overall behaviour and functioning.

SHAP is a method for explaining individual predictions based on Shapley values, which are optimal from the perspective of game theory. Shapley's values are a measure used in cooperative game theory to fairly allocate the value or contribution of each player to a specific game or problem. Shapley values are based on the idea that a player's value in a game depends on their contribution relative to the different possible coalitions or combinations of players. In other words, a player's value is calculated by considering all the possible ways they could have collaborated with the other players. To calculate Shapley's values, all possible permutations of players are considered, and the change in the game value when an additional player is added to the coalition is evaluated. These changes are averaged to obtain a fair measure of each player's value (Molnar 2022). In our context, the players are the different

features of the model, and the game value is the prediction outcome (the estimated success percentage of the attempt).

To calculate *SHAP* values, we selected the Python library *SHAP* (Lundberg 2018). Specifically, we employed the *explainer dashboard*, which builds upon *SHAP* to offer the option of visually representing SHAP values through dashboards. It is important to note that the importance of features in an individual prediction has a local interpretation. In other words, this importance is also influenced by the values of the remaining features. Therefore, in two different attempts where one feature holds the same value but the other features vary, they may have different levels of importance for the prediction. Thus, it is crucial not to attempt to explain the importance of a feature in isolation without considering the values of the other features, as these values also influence the significance of that particular feature.

4 | Results

4.1 | Performance Prediction Models: Comparison and Identification of the Best Model for Prediction

The first step after preprocessing was the configuration and evaluation of the AI algorithms through a ten-fold cross-validation for each time interval. The best results obtained by each algorithm and time interval in the cross-validations are summarised in Tables 5–7. The results show that RF achieved the best performance in every time interval, achieving a balanced accuracy of 0.76 in the first time interval (25%), 0.772 in the second time interval (50%), and 0.795 in the third time interval (75%). It is noteworthy that the algorithms achieved high performance even in the first time interval, and the performance improvement when moving to the second and third

TABLE 5 | Ten-fold cross-validation training results for prediction 1 (25th).

Model	Balanced accuracy	F1 score	MCC	Precision	Sensitivity	Specificity
AdaBoost	0.7253	0.8887	0.5069	0.8513	0.9324	0.5183
DT	0.7448	0.8857	0.5106	0.9103	0.5794	0.8637
KNN	0.6746	0.8731	0.4092	0.8253	0.9297	0.4195
MLP	0.7408	0.8858	0.5077	0.8628	0.9118	0.5697
RF	0.7598	0.8801	0.5175	0.8824	0.8793	0.6403
SGD	0.7230	0.7548	0.3897	0.9029	0.6611	0.7848
SVC	0.6964	0.8314	0.3690	0.8551	0.8112	0.5816

TABLE 6 | Ten-fold cross-validation training results for prediction 2 (50th).

Model	Balanced accuracy	F1 score	MCC	Precision	Sensitivity	Specificity
AdaBoost	0.7621	0.8782	0.5722	0.8337	0.9309	0.5932
DT	0.7433	0.8293	0.4874	0.8187	0.6680	0.8460
KNN	0.6995	0.8371	0.4311	0.7949	0.8866	0.5124
MLP	0.7653	0.8620	0.5493	0.8430	0.8853	0.6453
RF	0.7715	0.8698	0.5635	0.8526	0.8911	0.6518
SGD	0.7096	0.7854	0.4118	0.8340	0.7529	0.6663
SVC	0.7183	0.8557	0.4891	0.8129	0.9068	0.5299

TABLE 7 | Ten-fold cross-validation training results for prediction 3 (75th).

Model	Balanced accuracy	F1 score	MCC	Precision	Sensitivity	Specificity
AdaBoost	0.7789	0.8261	0.5659	0.7908	0.8716	0.6863
DT	0.7565	0.7787	0.4981	0.7749	0.7380	0.7873
KNN	0.7217	0.7897	0.4604	0.7300	0.8676	0.5759
MLP	0.7885	0.8401	0.5985	0.7980	0.8921	0.6849
RF	0.7954	0.8383	0.6001	0.8090	0.8737	0.7172
SGD	0.7035	0.7271	0.4164	0.7394	0.7663	0.6407
SVC	0.7365	0.8087	0.5009	0.7488	0.8840	0.5889

time intervals is not particularly high. The main reasons are that the user and puzzle features contain relevant information for predicting student performance, and the attempt features quickly capture the student's skills.

Once we selected RF as the best algorithm, we assessed the generalisation power of the three RF models (one for each time interval) in the test set. The test set contained unseen data not employed during the models' training, configuration and selection. Table 8 shows the balanced accuracy achieved by RF in the test set of each time interval. The performance of RF was even better than in the cross-validation, showing a good generalisation power of the three models.

Based on the results mentioned above, we chose RF to be applied in the use case that we will present afterward. However, RF is a 'black box' algorithm. Therefore, our models are not inherently interpretable, unlike other algorithms such as KNN or decision trees. To improve the interpretability of the models, we applied the *SHAP* method, which allows us to explain individual predictions in the use case. Moreover, this method can also be useful for identifying the most relevant factors contributing to students' success when solving puzzles. To achieve this, we rely on the importance of features calculated as the average absolute value of the Shapley values obtained by each feature in different individual predictions. We can see the importance of each feature in each prediction in Figure 3. **Puzzle features** are shown in red, **user**

TABLE 8 | Test results.

Model	Balanced accuracy	F1 score	MCC	Precision	Sensitivity	Specificity
Prediction 1 (25th)	0.7721	0.8445	0.5439	0.8449	0.8441	0.7
Prediction 2 (50th)	0.7918	0.8288	0.5912	0.8026	0.8567	0.7268
Prediction 3 (75th)	0.7928	0.8020	0.5877	0.7723	0.8340	0.7516

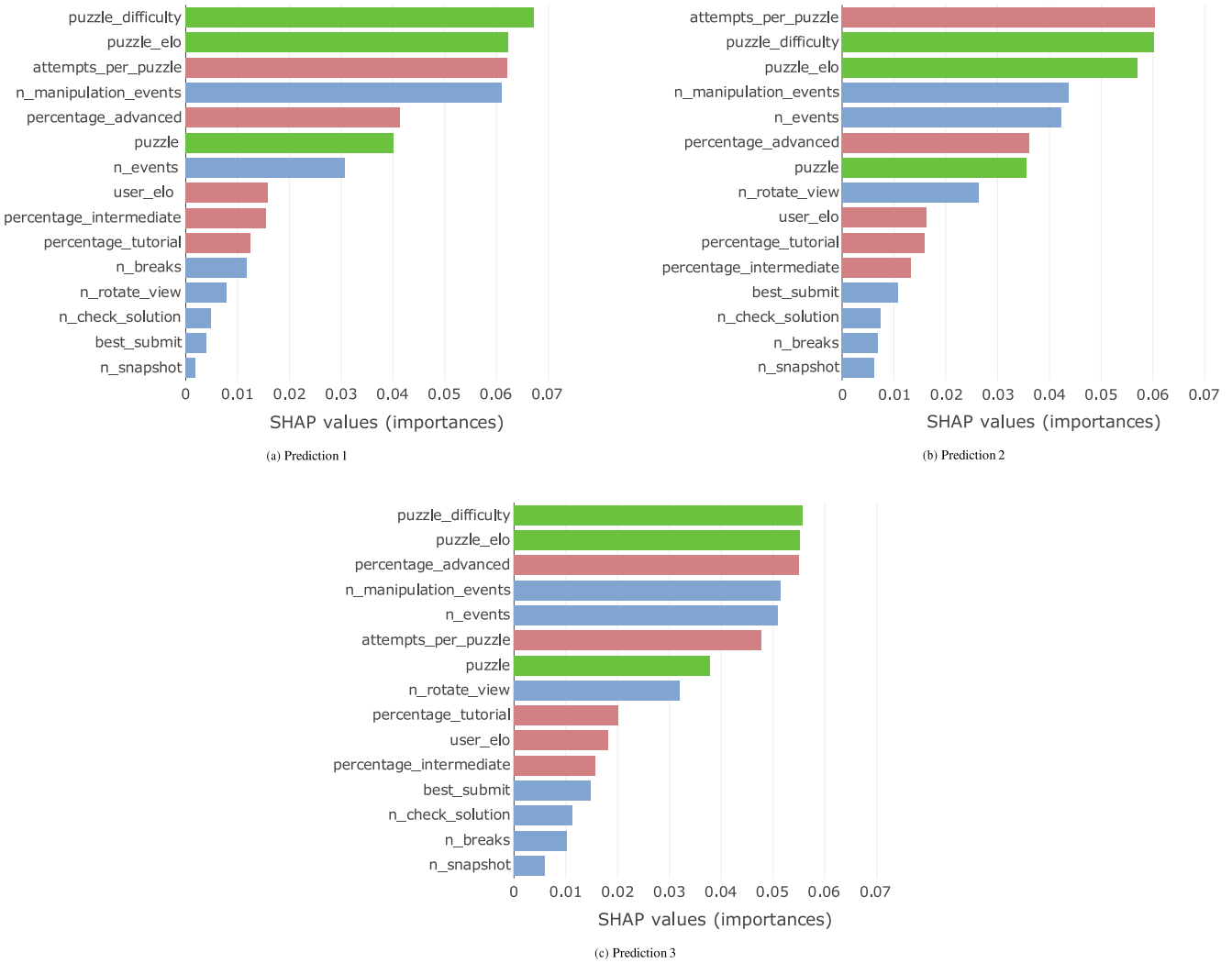

FIGURE 3 | Importance of each feature for each prediction.

features in green and **attempt features** in blue. Although there are a few exceptions, the relevance of most features remains relatively stable across predictions.

We observe that the most crucial features across all predictions are puzzle-related, specifically `puzzle_difficulty` and `puzzle_elo`. This implies that the decisive factor in determining whether a user will solve a puzzle is the puzzle's difficulty. Moving to **user features**, both the percentage of advanced puzzles completed (`percentage_advanced`) and the number of attempts per puzzle (`attempts_per_puzzle`) hold significant importance. In contrast, the percentage of intermediate puzzles completed (`percentage_intermediate`) and the user Elo (`user_elo`) show low relevance. This suggests that efficient students who have successfully tackled advanced puzzles and made few mistakes are more likely to solve the puzzle they currently face. Finally, regarding **attempt features**, we note that the number of events (`n_events` and `n_manipulation_events`) plays a crucial role in prediction. Therefore, in order to successfully solve a puzzle, it is essential that the user is proactive, actively creating and manipulating figures.

Finally, we can take a look at the remaining **attempt features**: `n_check_solution`, `n_rotate_view`, `n_breaks`, `best_submit` and `n_snapshot`. All these features have low relevance in the prediction, except for the number of camera rotations, which gains relevance in the third prediction. Considering this, it is not crucial whether a student takes prolonged pauses during puzzle-solving, the number of screenshots taken, or the best solution achieved up to that moment. However, we can interpret that performing more camera rotations as the solving process advances helps users find a solution when stuck. This action offers new perspectives on the scenario, which is a crucial aspect in spatially related geometric problems.

4.2 | Use Case: Supporting Individual Students in the Classroom

For this case study, we envision a classroom scenario in which a teacher is using the tool and has access to a dashboard to monitor struggling students. This dashboard can identify students with low probabilities of completing a level, enabling the teacher to prioritise them for JIT support and assistance in completing the task. To do so, the teacher can analyse the individual predictions of the model displayed on the dashboard to better understand its functioning. We will present two examples from our dataset, one where the model corresponding to the second prediction correctly predicts a failed attempt and another where it correctly predicts the success of an attempt. Additionally, we will analyse the obtained *SHAP* values for each feature to study how the teacher can interpret these values and assist students appropriately.

Regarding the first student, Table 9 shows the features associated with this student's second attempt at this level, along with the obtained *SHAP* values for each feature. The user who made this attempt has completed 88.89% of the tutorial puzzles (`percentage_tutorial`), none of the intermediate puzzles(`percentage_intermediate`), and 7.89% of the advanced puzzles(`percentage_advanced`). The student has a `user_elo` value of 0.08 and an average of 2.33 attempts per completed puzzle (`attempts_per_puzzle`).

TABLE 9 | Features and *SHAP* values of the first student's attempt.

Feature	Value	SHAP
percentage_tutorial	88.89	−0.0161
percentage_intermediate	0.00	0.0157
percentage_advanced	7.69	0.0019
attempts_per_puzzle	2.33	−0.0648
user_elo	0.08	−0.0014
puzzle_elo	3.00	−0.126
puzzle_difficulty	1.00	−0.1265
n_events	17.00	0.0176
n_check_solution	0.00	0.0056
best_submit	0.00	−0.0064
n_breaks	0.00	−0.0008
n_manipulation_events	6.00	0.0228
n_snapshot	0.00	−0.0034
n_rotate_view	0.00	−0.0193
puzzle	Bear market	−0.0777
completed	0.0	

The puzzle that the student is trying to solve is 'Bear market', which is the most difficult puzzle in the entire game according to the `puzzle_elo` and `puzzle_difficulty` features. The user has performed 17 events (`n_events`), out of which six are manipulation events (`n_manipulation_events`). The value of `n_breaks`, `n_check_solution`, `n_rotate_view` and `n_snapshot` features is zero. This means that there have been no periods of inactivity, no attempts to check the puzzle solution, and no camera rotations or screenshots taken. The last row of the table indicates that the user did not manage to complete the puzzle in that particular attempt.

In Figure 4, the *SHAP* values and how they contribute to the final prediction are visually presented. The first bar indicates the model's prediction result without considering any features. As expected, this bar indicates 50% since, without any features, the model cannot determine whether to predict success or failure. Remember that if the final prediction percentage is higher than 50%, it will predict a successful resolution (one), and otherwise, it predicts failure (zero).

The following bars correspond to the obtained *SHAP* values (Table 9). Starting from the initial prediction percentage (50%), the corresponding *SHAP* values are added or subtracted until reaching the final prediction percentage (e.g., a *SHAP* value of 0.025 translates to a 2.5% increase in the final prediction). Green bars represent positive contributions, while red bars represent negative contributions. The most relevant

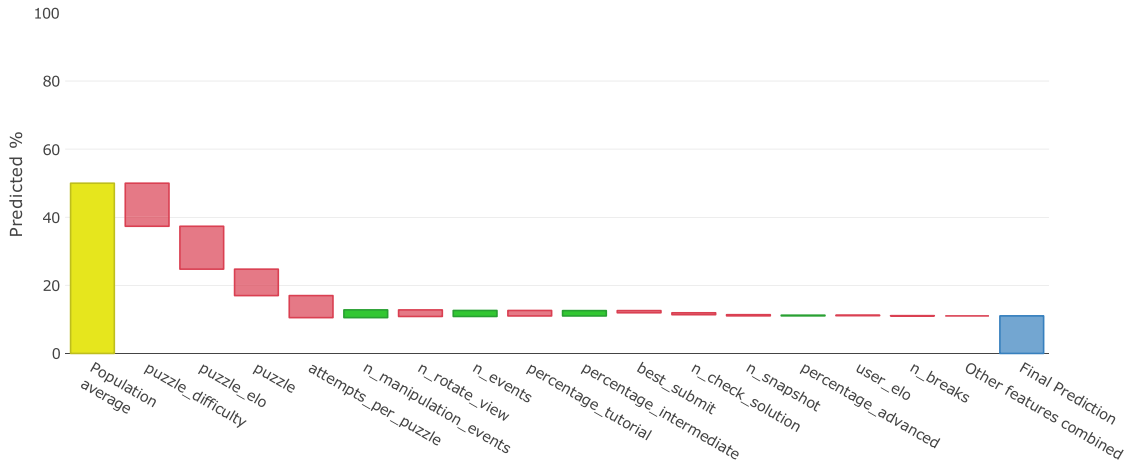

FIGURE 4 | SHAP values of the first student's attempt.

features in this prediction are related to the puzzle. Given that it is the most challenging puzzle in the game, the contribution of each of these features is highly negative to the final prediction. Additionally, the user has a high average number of attempts per completed puzzle (`attempts_per_puzzle`), which also has a significant negative impact on the prediction. Consequently, the model predicts a success percentage of only 11.03%, accurately predicting the user's failure to solve the puzzle.

Given this information provided by the dashboard, the teacher has the option to suggest that students start tackling easier puzzles since this student has begun solving the most difficult puzzle in the game. Thus, the teacher may advise them to concentrate on solving simpler puzzles first, such as the intermediate ones, especially if they have not completed any of them.

Now, we analyse an attempt from a different student. In Table 10, we can observe the features corresponding to the first attempt of this student in this particular puzzle. The user who played this attempt completed 55.6% of the tutorial puzzles and none of the intermediate or advanced puzzles. The student has a `user_elo` value of 0.0 (Elo does not consider tutorial puzzles) and an average of 1.0 attempt per completed puzzle, meaning that the user has successfully completed all the previous puzzles in the first attempt. Regarding the puzzle, in this case, it is '6. Stretch a ramp', which has a difficulty value of 0.11 and an Elo value of 0.42, making it relatively easy to solve. Up to that moment, the user has performed a total of 31 events (of which 15 were manipulation events), has had no periods of inactivity and neither rotated the camera nor took screenshots. Finally, the last row of the table indicates that the user successfully solved the puzzle in this attempt.

In Figure 5, the *SHAP* values and their contributions to the final prediction are visually presented. Starting from the initial prediction percentage (50%), we observe that the feature contributing the most is the number of attempts per completed puzzle. Since this user has completed all attempted puzzles on the first try, the model assigns significant relevance to this feature. Subsequently, puzzle-related features also contribute significantly to the decision, given the low difficulty of this puzzle. Finally, it is worth noting the relevance of manipulation events.

TABLE 10 | Features and *SHAP* values of the second student's attempt.

Feature	Value	SHAP
percentage_tutorial	55.56	0.0205
percentage_intermediate	0.00	0.0163
percentage_advanced	0.00	0.03
attempts_per_puzzle	1.00	0.0893
user_elo	0.00	0.0189
puzzle_elo	0.42	0.0725
puzzle_difficulty	0.11	0.0713
n_events	31.00	0.0364
n_check_solution	0.00	0.0069
best_submit	0.00	−0.0097
n_breaks	0.00	0.0073
n_manipulation_events	15.00	0.0497
n_snapshot	0.00	−0.0003
n_rotate_view	0.00	−0.0282
puzzle	6 Stretch a ramp	0.0414
completed	1.0	

The user has extensively created and manipulated figures, a clear indication of being close to a puzzle solution, thus contributing positively to the model's prediction. In the end, the model predicts that the student will solve the puzzle with a probability of 92.3%. Given this scenario, the teacher does not need to assist the student due to the context provided by the features and the high probability of solving the puzzle.

5 | Discussion

The analysis and interpretation of data generated by SGs can provide valuable information for learners and instructors in educational settings. For example, instructors can follow a student's

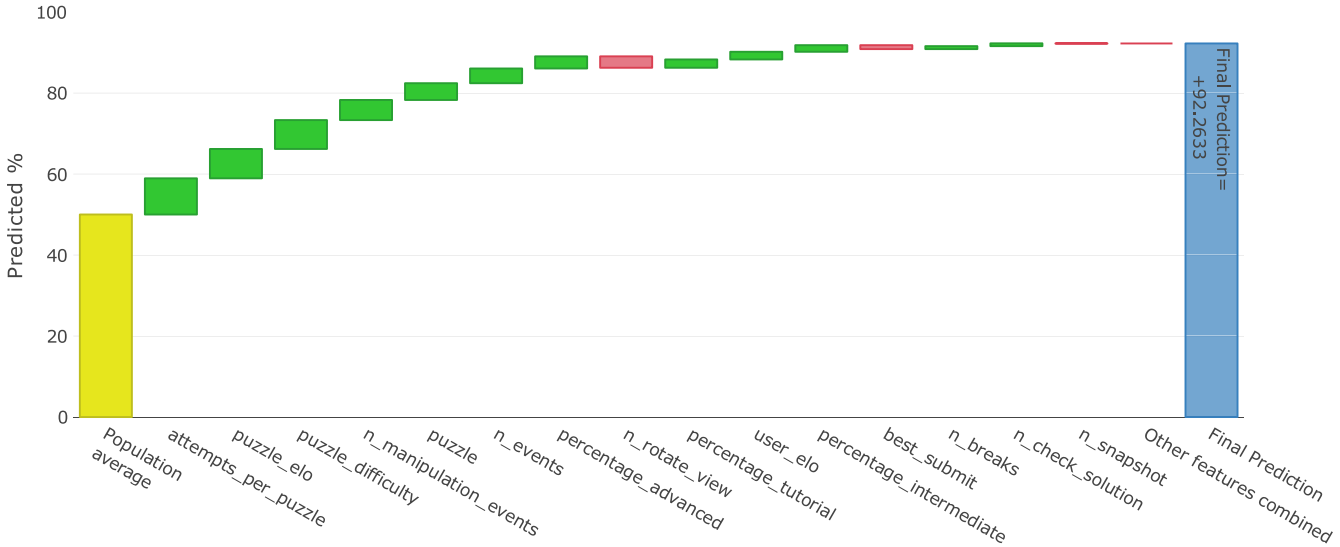

FIGURE 5 | SHAP values of the second student's attempt.

progression in real-time while playing and take action on any identified learning problems (Serrano-Laguna et al. 2017). This data analysis would enable the application of JIT interventions in classrooms. By continuously monitoring and interpreting students' interactions, game analytics allow instructors to identify learning gaps and challenges as they emerge. However, the analyses typically used in GBA studies are quite simple, and there is still a present challenge in developing more sophisticated and robust methods for leveraging the vast amount of data generated by SGs (Gomez, Ruipérez-Valiente, and Clemente 2023).

In this regard, AI models can be used to predict future performance based on current learning behaviours, adding a novel dimension to the personalisation and adaptation of GBAs (Kickmeier-Rust 2018). In this research, we developed a predictive approach that integrates AI as a crucial component in analysing this game data, taking into account several features derived from students' interactions with the game and allowing us to assess and predict students' performance while playing. These predictions can be useful in many ways. For instance, they enable instructors to intervene in the classroom when necessary, identifying students who are experiencing difficulties and helping them in real-time. Moreover, this information could also be used as input in adaptive platforms to adjust the game difficulty based on the predictions, provide hints and adjust the learning experience to each student's needs.

However, for instructors to provide this personalised learning approach, they need to understand the information generated by the AI models. XAI plays a crucial role in facilitating the socio-cultural process of learning, where interactions between teachers and students are fundamental in guiding learners through zones of proximal development and providing personalised support to students facing difficulties (Khosravi et al. 2022). Incorporating XAI methods further enhances the educational benefits of our proposal. XAI techniques facilitate the comprehension and explanation of how our AI models make predictions, and this transparency enables instructors to understand the reasoning the model followed to make the prediction and act accordingly.

Moreover, an important aspect of AI systems is to ensure user trust. Explainability gives users confidence that AI systems work well, helps developers understand why a system works a certain way, and safeguards against bias (Shin 2021). In education, the need for explanation arises since educators must be accountable to students, parents, and the government. Explanation is crucial when providing individual feedback to students, offering teachers diagnostic feedback to identify areas where a class of students needs increased focus, and during parental consultations to help them support their child's learning (Khosravi et al. 2022). By incorporating XAI techniques on top of our models, we increase the transparency of our predictive models in educational contexts, allowing instructors to understand the reasoning behind the predictions and building confidence in the potential use of SGs in the classroom.

6 | Conclusions

This research aimed to develop an XAI model for predicting students' performance in real-time while playing *Shadowspect*, a geometry SG. The RF predictive model developed in this study demonstrates a promising accuracy in anticipating students' task completion, achieving a balanced accuracy result of 77.21% in making early predictions after a short playtime has elapsed. Moreover, we ensured that the model predictions are fully explainable by taking into account both intrinsic and extrinsic explainability options. This way, our work provides a comprehensive framework for interpretable models, enabling a better understanding of the AI model predictions and facilitating informed decision-making in educational contexts. With this research, we aim to contribute to the educational field by providing a powerful and understandable tool that supports personalised learning experiences and effectively integrates SGs into educational settings.

This work has some limitations. First, our ML models were built using a specific set of log data from *Shadowspect*, which may not fully capture the diversity of behaviours present in other contexts and SGs. Expanding the feature set to include

data from other SGs could improve the models' applicability across different games and enhance predictive accuracy. Moreover, although our interpretable models are useful for understanding the model's decisions, the output might still be complex for non-technical users. Thus, we plan to refine the model's interpretability to ensure that even users with little technical expertise can benefit from the insights provided by the models. Furthermore, while we acknowledge the effectiveness of the ML models developed, the use of more complex techniques, such as DL, could potentially enhance the predictive power and robustness of our approach. Another limitation is that our data comes from a partially controlled classroom setting. As part of our future work, we intend to conduct case studies and experiments in various real-world educational settings to better understand the adaptability and generalisability of our approach. By addressing these limitations, we aim to make our model a valuable tool for personalised learning and interventions.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Research data are not shared.

References

Abeyrathna, D., S. Vadla, V. Bommanapally, M. Subramaniam, P. Chundi, and A. Parakh. 2019. "Analyzing and Predicting Player Performance in a Quantum Cryptography Serious Game." In *Games and Learning Alliance. GALA 2018. Lecture Notes in Computer Science*, edited by M. Gentile, M. Allegra, and H. Söbke, vol. 11385. Cham, Switzerland: Springer.

Al Qallawi, S., and M. Raghavan. 2022. "A Review of Online Reactions to Game-Based Assessment Mobile Applications." *International Journal of Selection and Assessment* 30, no. 1: 14–26.

Alamri, R., and B. Alharbi. 2021. "Explainable Student Performance Prediction Models: A Systematic Review." *IEEE Access* 9: 33132–33143.

Alonso-Fernández, C., I. Martínez-Ortiz, R. Caballero, M. Freire, and B. Fernández-Manjón. 2020. "Predicting Students' Knowledge After Playing a Serious Game Based on Learning Analytics Data: A Case Study." *Journal of Computer Assisted Learning* 36, no. 3: 350–358.

Anderson, A., and E. Wood. 2009. "Implementing Technology in the Classroom: Assessing Teachers' Needs Through the Use of a Just-In-Time Support System." In *Society for Information Technology & Teacher Education International Conference*, 3369–3372. Charleston, SC, USA: Association for the Advancement of Computing in Education (AACE).

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, et al. 2020. "Explainable Artificial Intelligence (Xai): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible Ai." *Information Fusion* 58: 82–115.

Bado, N. 2022. "Game-Based Learning Pedagogy: A Review of the Literature." *Interactive Learning Environments* 30, no. 5: 936–948.

Bellotti, F., R. Berta, and A. De Gloria. 2010. "Designing Effective Serious Games: Opportunities and Challenges for Research." *International Journal of Emerging Technologies in Learning (iJET)* 5, no. 2010: 22–35.

Berger, F., and W. Müller. 2021. "Back to Basics: Explainable Ai for Adaptive Serious Games." In *Serious Games. JCSG 2021. Lecture Notes in Computer Science*, edited by B. Fletcher, M. Ma, S. Göbel, J. B.Hauge, and T. Marsh, vol. 12945. Cham, Switzerland: Springer.

Chitti, M., P. Chitti, and M. Jayabalan. 2020. "Need for Interpretable Student Performance Prediction." In *2020 13th International Conference on Developments in Esystems Engineering (Dese)*, 269–272. IEEE: Liverpool, United Kingdom.

de Klerk, S., and P. M. Kato. 2017. "The Future Value of Serious Games for Assessment: Where Do We Go Now?" *Journal of Applied Testing Technology* 18, no. S1: 32–37.

Elo, A. 2008. *The Rating of Chessplayers: Past and Present*. Bronx, NY: Ishi Press International.

Farrow, R. 2023. "The Possibilities and Limits of Xai in Education: A Socio-Technical Perspective." *Learning, Media and Technology* 48, no. 2: 266–279.

Gomez, M. J., J. A. Ruipérez-Valiente, and F. J. G. Clemente. 2023. "A Systematic Literature Review of Game-Based Assessment Studies: Trends and Challenges." *IEEE Transactions on Learning Technologies* 16, no. 4: 500–515.

Gunning, D. 2017. "Explainable Artificial Intelligence (XAI)." *Defense Advanced Research Projects Agency (DARPA), nd Web* 2, no. 2: 1.

Hooshyar, D., N. El Mawas, M. Milrad, and Y. Yang. 2023. "Modeling Learners to Early Predict Their Performance in Educational Computer Games." *IEEE Access* 11: 20399–20417.

Illanas Vila, A., J. Calvo Ferrer, F. Gallego Durán, and F. Llorens Largo. 2013. "Predicting Student Performance in Translating Foreign Languages With a Serious Game." In *INTED2013 Proceedings*, 52–59. IATED: Valencia, Spain.

Jang, Y., S. Choi, H. Jung, and H. Kim. 2022. "Practical Early Prediction of Students' Performance Using Machine Learning and Explainable AI." *Education and Information Technologies* 27, no. 9: 12855–12889.

Khosravi, H., S. B. Shum, G. Chen, et al. 2022. "Explainable Artificial Intelligence in Education." *Computers and Education: Artificial Intelligence* 3: 100074.

Kickmeier-Rust, M. D. 2018. "Predicting Learning Performance in Serious Games." In *Serious Games: 4th Joint International Conference, JCSG 2018, Darmstadt, Germany, November 7–8, 2018, Proceedings*, vol. 4, 133–144. Cham: Springer International Publishing.

Laamarti, F., M. Eid, and A. El Saddik. 2014. "An overview of serious games." *International Journal of Computer Games Technology* 2014: 1–15.

Larson, K. 2020. "Serious Games and Gamification in the Corporate Training Environment: A Literature Review." *TechTrends* 64, no. 2: 319–328.

Lee, J.-E., A. Jindal, S. N. Patki, A. Gurung, R. Norum, and E. Ottmar. 2023. "A Comparison of Machine Learning Algorithms for Predicting Student Performance in an Online Mathematics Game." *Interactive Learning Environments* 32, no. 9: 5302–5316.

Loh, C. S., Y. Sheng, and I.-H. Li. 2015. "Predicting Expert–Novice Performance as Serious Games Analytics With Objective-Oriented and Navigational Action Sequences." *Computers in Human Behavior* 49: 147–155.

Lundberg, S. 2018. "Shap Documentation." https://shap.readthedocs.io/en/latest/index.html.

Marín-Morales, J., L. A. Carrasco-Ribelles, M. Alcañiz, and I. A. C. Giglioli. 2021. "Applying Machine Learning to a Virtual Serious Game for Neuropsychological Assessment." In *2021 IEEE Global Engineering Education Conference (Educon)*, 946–949. IEEE: Vienna, Austria.

Molnar, C. 2022. *Interpretable Machine Learning*. 2nd ed. Victoria, Canada: Leanpub. https://christophm.github.io/interpretable-ml-book.

Papanastasiou, G., A. Drigas, C. Skianis, and M. D. Lytras. 2017. "Serious Games in k-12 Education: Benefits and Impacts on Students With Attention, Memory and Developmental Disabilities." *Program* 51, no. 4: 424–440.

Preece, A. D., D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty. 2018. "Stakeholders in Explainable AI." *CoRR*, *abs/1810.00184*.

Rastrollo-Guerrero, J. L., J. A. Gómez-Pulido, and A. Durán-Domínguez. 2020. "Analyzing and Predicting students' Performance by Means of Machine Learning: A Review." *Applied Sciences* 10, no. 3: 1042.

Ruipérez-Valient, J. A., Y. J. Kim, R. S. Baker, P. A. Martínez, and G. C. Lin. 2022. "The Affordances of Multivariate Elo-Based Learner Modeling in Game-Based Assessment." *IEEE Transactions on Learning Technologies* 16, no. 2: 152–165.

Ruiperez-Valiente, J. A., M. Gaydos, L. Rosenheck, Y. J. Kim, and E. Klopfer. 2020. "Patterns of Engagement in an Educational Massively Multiplayer Online Game: A Multidimensional View." *IEEE Transactions on Learning Technologies* 13, no. 4: 648–661.

Serrano-Laguna, Á., I. Martínez-Ortiz, J. Haag, D. Regan, A. Johnson, and B. Fernández-Manjón. 2017. "Applying Standards to Systematize Learning Analytics in Serious Games." *Computer Standards & Interfaces* 50: 116–123.

Shin, D. 2021. "The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable Ai." *International Journal of Human-Computer Studies* 146: 102551.

Smith, S. P., K. Blackmore, and K. Nesbitt. 2015. "A Meta-Analysis of Data Collection in Serious Games Research." In *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, 31–55. Cham: Springer International Publishing.

Swarz, J., A. Ousley, A. Magro, et al. 2010. "Cancerspace: A Simulation-Based Game for Improving Cancer-Screening Rates." *IEEE Computer Graphics and Applications* 30, no. 1: 90–94. https://doi.org/10.1109/MCG.2010.4.

Tao, J., Y. Xiong, S. Zhao, et al. 2020. "Xai-Driven Explainable Multi-View Game Cheating Detection." In *2020 IEEE Conference on Games (Cog)*, 144–151. IEEE: Osaka, Japan.

Wiemeyer, J., and A. Kliem. 2012. "Serious Games in Prevention and Rehabilitation—A New Panacea for Elderly People?" *European Review of Aging and Physical Activity* 9, no. 1: 41–50.

Yuhana, U. L., R. G. Mangowal, S. Rochimah, E. M. Yuniarno, and M. H. Purnomo. 2017. "Predicting Math Performance of Children With Special Needs Based on Serious Game." In *2017 IEEE 5th International Conference on Serious Games and Applications for Health (Segah)*, 1–5. Perth, WA, Australia: IEEE.

Zawacki-Richter, O., V. I. Marín, M. Bond, and F. Gouverneur. 2019. "Systematic Review of Research on Artificial Intelligence Applications in Higher Education–Where Are the Educators?" *International Journal of Educational Technology in Higher Education* 16, no. 1: 1–27.