A Scalable, Flexible, and Interpretable Analytic Pipeline for Stealth Assessment in Complex Digital Game-Based Learning Environments: Towards Generalizability

Wenyi Lu
University of Missouri
Columbia, MO
65211, USA
wldh6@umsystem.edu

James Laffey
University of Missouri
Columbia, MO
65211, USA
LaffeyJ@missouri.edu

Troy D. Sadler
The University of
North Carolina at
Chapel Hill
Chapel Hill, NC
27599, USA
tsadler@unc.edu

Joseph Griffin
University of Missouri
Columbia, MO
65211, USA
griffinjg@missouri.edu

Sean P. Goggins
University of Missouri
Columbia, MO
65211, USA
gogginss@missouri.edu

The rapid advancement of technology necessitates innovative educational tools and curricula that empower learners to acquire and apply new knowledge and skills, particularly for complex, real-world problem-solving. Digital Game-Based Learning (DGBL) has emerged as a promising approach to engage students in meaningful learning experiences. However, one major challenge for DGBL adoption in formal education is the effective assessment of learners' performance aligned with specific educational standards, such as the Next Generation Science Standards (NGSS). This study addresses this challenge by proposing and evaluating a novel stealth assessment (SA) pipeline that leverages educational data mining techniques to enhance the generalizability and scalability of learning assessments across various DGBL contexts, while maintaining model interpretability and improving the flexibility of model selection. Our proposed analytical pipeline integrates both machine-learned and expert-crafted features to predict multiple learning outcomes, content knowledge, and scientific argumentation skills. The pipeline offers several innovations: (1) it captures intricate in-game behaviors and decision-making strategies; (2) it employs a three-layered unsupervised learning approach to reduce dimensionality and identify critical features; and (3) it provides a flexible framework by combining both in-game and learning progress data. We validate this pipeline within a 3D narrative DGBL environment, Mission HydroSci (MHS), demonstrating its utility in accurately assessing learning outcomes across multiple game contexts (units). Moreover, by employing Accumulated Local Effects (ALE) plots, this study interprets the black-box models' results, offering actionable insights into game design and pedagogical arrangements. Our findings reveal unexpected relationships between in-game performance and post-game learning outcomes, leading to recommendations for future DGBL design improvements. This study advances educational data mining by providing a scalable, flexible and interpretable framework for embedding SA into DGBL environments, thus extending the reach of data-driven learning assessments in educational game contexts. Future research will further explore the applicability and limitations of this pipeline across diverse educational settings. Codes and sample datasets can be found at `https://github.com/augurlabs/2025-Lu-Et-Al`

Keywords: stealth assessment, digital game-based learning, evidence-centered design, educational data mining, machine learning, learning analytics, unsupervised learning, dimension reduction, latent variable learning, ensemble learning, accumulated local effects plots

1. Introduction

As technology advances rapidly, the importance of individuals being able to quickly acquire new knowledge and skills, especially those related to complex real-world problem solving, is increasingly crucial (Bahrami et al., 2023). This demand necessitates cutting-edge curricula and instructional tools that equip students with the necessary competencies, particularly the capability for sustainable learning, to confront the uncertainties of present and future society. As stated in the quote, “If we teach today as we taught yesterday, we are destroying the future of our children” (Dewey, 1974), the need for educational innovations to help individuals keep pace with a constantly evolving world cannot be overstated.

Educators highlight the importance of advancing learning technologies that help motivate and engage learners in applying their newly acquired knowledge and skills in new situations (Gaikwad, 2022; Steinemann, 2003). digital game-based learning (DGBL) aligns well with these requirements, given its inherent features and potential to engage (Eseryel et al., 2014; Yang, 2012). Its scaffolded task design and captivating storylines motivate learners to explore and solve problems at their own pace (Sun et al., 2011; Rowe et al., 2010). Additionally, learners can form emotional connections with their in-game characters, which increases engagement (Plass et al., 2020). Simulations of real-world game environments provide rich opportunities for learners to acquire and apply the knowledge they can transfer to real-world situations (Barab et al., 2010). Games hold the potential to offer real-time formative feedback and assessment based on learners’ in-game behaviors, helping them adjust their learning strategies (Leonardou et al., 2020).

Previous research has demonstrated the efficacy of DGBL in teaching STEM subjects (Wang et al., 2022), problem-solving skills (Miladinovic et al., 2023; Liu and Israel, 2022), computational thinking (Lu et al., 2023), creativity (Nie et al., 2014), and language acquisition as well as social development (Darvenkumar and Devi, 2022). Despite these positive outcomes, the widespread adoption of DGBL as an instructional tool in educational settings continues to face challenges. One of the most significant challenges is that in-game assessments are aligned primarily with a particular game’s content, complicating accurate measurement of learners’ performance related to the targeted learning outcomes, and making it harder to measure whether learners meet established educational standards, such as the Next Generation Science Standards (NGSS) (Sanchez and Lee, 2022; Nguyen et al., 2020).

To address DGBL assessment challenge, researchers have proposed utilizing external assessments, such as pre-and post-tests, to evaluate the targeted learning objectives independent of the progress made by learners during gameplay (Caballero-Hernández et al., 2024). Furthermore, conducting assessments for each learning objective within the game is recommended using formats such as pop-up text boxes with multiple-choice options. However, when not well-designed, these methods can significantly hinder learners' gaming immersion and engagement because of sudden interruption and test anxiety (Steinmaurer et al., 2021; Frommel et al., 2015; Bellotti et al., 2013).

1.1. Stealth Assessment

To address DGBL assessment needs, stealth assessment (SA), which is seamlessly integrated into the game design, unobtrusively measures the performance and learning outcomes of learners at various game stages based on data-driven approaches (Shute et al., 2009). Motivated by the potential advantages of SA in DGBL, numerous studies demonstrate promising research outcomes using this method (Shute et al., 2021; Henderson et al., 2020; Min et al., 2019; Yang et al., 2021; Gris and Bengtson, 2021). One framework is implementing SA as part of an evidence-centered design (ECD) approach, which is a systematic way for embedding assessments to evaluate learning objectives based on evidence from learners' behaviors within DGBL (Mislevy et al., 2003; Shute et al., 2009). Building on ECD, Shute and colleagues implement SAs in various DGBL environments to demonstrate their effectiveness in measuring problem-solving (Shute et al., 2016), mathematics (Smith et al., 2019), conscientiousness (Moore and Shute, 2017), and creativity (Shute and Rahimi, 2021) using Bayesian networks (BN). BN, a "white-box" machine learning model, presents a clear tree-based visualization of the relationship between predictors (features) and learning outcomes. By analyzing the BN structure, researchers can determine the optimal combination of feature values to achieve the best learning outcome. However, constructing a robust BN is labor-intensive and time-consuming. Training requires a lot of data, leading to potential overfitting issues and difficulties in applying models generated in one DGBL environment within another DGBL environment (Georgiadis et al., 2019). Additionally, the validation of BN predictions against external peer-reviewed assessments is not always guaranteed. For example, Shute et al. (2016) used two external assessments - Raven's Progressive Matrices and MicroDYN - to validate their BN predictions for problem-solving skills. While external assessments validated the BN's estimates for overall problem-solving skills and some facets, certain facets were not fully aligned with the external assessments. This is primarily due to the limitations in capturing complex, non-linear relationships and the context-specific nature of BN-based competency models.

Recent studies in SA within DGBL environments have explored integrating machine learning algorithms to streamline feature engineering, address sparse data challenges, and predict learning outcomes (Gupta et al., 2021; Min et al., 2019; Henderson et al., 2022). For instance, the DeepStealth framework, based on the ECD approach, leverages raw game logs and deep-learning techniques to reduce reliance on manual feature engineering and expand applicability across diverse contexts (Min et al., 2019). While these advancements offer promise, challenges such as model interpretability and generalizability to complex game mechanics remain (Akram et al., 2018). This underscores the need for further research into scalable and interpretable SA frameworks.

1.2. Seeking Multi-Game Utility from Stealth Assessment

In pursuit of a more generalizable approach to SA within DGBL contexts, Georgiadis et al. (2019) proposed a design concept model. Using a Realising Applied Gaming Ecosystem (RAGE) architecture of client-side applied gaming components, they developed their SA prototype to make it useful in any DGBL context. They verified their SA prototype's technical feasibility by generating simulation data sets with different conditions, such as sample size, data type, and probability distribution based on the xAPI standard (Georgiadis et al., 2020). Then, they trained machine learning models to predict competency achievement using different combinations of data facets. While the prediction accuracy of their models reached 90% accuracy in all considered conditions, they note the need for extensive empirical evaluation of their SA approach for DGBL to verify its feasibility for use across DGBLs.

The research presented in this paper describes and evaluates an SA educational data mining pipeline that advances knowledge in the field toward generalizability. It addresses the following deficits: 1) Scalability problems. We operationalize our pipeline for different segments or contexts (each unit and the whole game) of a complex DGBL environment, which allows students to explore freely and solve problems mimicking those in real-world's. 2) Manual feature generation processes are labor-intensive, time-consuming, tied to a specific DGBL context, and challenging to transfer to other DGBLs. The pipeline described is more automated than others previously presented and transferrable to other DGBL environments. 3) Predicting only one aspect of learning subject matter. Our proposed SA measures a player's competency in not only content knowledge (in this case, water science), but scientific argumentation skills as well. 4) Limited flexibility in model selection. Our proposed pipeline facilitates greater flexibility in contrast to conventional approaches that primarily rely on BNs. While BNs are inherently interpretable, they often constrain the flexibility needed for diverse assessment contexts. By integrating both machine-learned and expert-crafted features, our pipeline maintains interpretability by using methods like accumulated local effects (ALE) plots, which provide actionable insights into model predictions. This approach enables educators and designers to better understand the relationships between in-game behaviors and learning outcomes, thereby supporting informed decision-making in pedagogical practices and game design refinements.

The resulting analytical pipeline provides a complete assessment framework for DGBLs, including complex game mechanics, dynamic game worlds to interact with, and ill-defined problems for learners to solve without binding SA to specific in-game action formats (e.g., selecting a choice within a dialogue). In this paper, we advance prior work that, using extensive empirical data from a DGBL system, provides a framework for systematically making SA a component of DGBL construction practice. We extend our previous studies and propose a novel analytical pipeline using SA in Mission HydroSci (MHS) (Laffey et al., 2019; Laffey et al., 2019). This 10-hour first-person 3D narrative adventure teaches middle school students water science and scientific argumentation in ways that fit the NGSS, which is central to our content knowledge assessments.

1.3. Research Questions

Specifically, this paper investigates three research questions to systematically contribute a more generalizable approach to SA in DGBL:

1. Research Question 1 (RQ1): What distinct elements are encapsulated within the overarching analytic pipeline?
2. Research Question 2 (RQ2): How effective is this pipeline across various contexts? Specifically:

(a) How accurate is the assessment across different MHS units?

(b) What is the overall assessment accuracy when considering MHS as a whole?

(c) How does assessment accuracy vary across distinct subjects within MHS?

3. Research Question 3 (RQ3): What methods can help interpret the black-box computational models, and what insights can be drawn from their results?

2. Related Work

2.1. Machine Learning and Artificial Intelligence Application in Digital Game-Based Learning

Incorporating artificial intelligence (AI) and machine learning techniques in educational domains has attracted significant interest among researchers and practitioners. Innovative findings indicate that AI-driven applications and platforms can considerably boost both the effectiveness and efficiency of learning and teaching processes, attributed to their capabilities for precise performance measurement, timely formative feedback, and personalized learning experiences. These advancements span a range of domains, such as learning technology and platform design (Mousavinasab et al., 2021; Luckin and Cukurova, 2019; Alam, 2021), assessment and evaluation (Hooshyar et al., 2016; Aluthman, 2016), competence and skill development (Lin and Chen, 2020; Sakulkueakulsuk et al., 2018; Ciolacu et al., 2018; Chopade et al., 2019), learning analytics and student behavior measurement (Blikstein and Worsley, 2016; Sharma et al., 2019; Yang et al., 2021; Doleck et al., 2020), and instructional design and optimization (Conati et al., 2018; Rosé et al., 2019; Eliseyev and Aksenova, 2019; Peng et al., 2022).

In particular, the fusion of AI and machine learning technologies with DGBL environments has also generated substantial interest. DGBL contexts offer captivating settings and immersive narratives stimulating interest-driven motivation and cognitive curiosity (Jackson et al., 2018; Naul and Liu, 2020). Moreover, these environments provide multifaceted interactivity, intricate reward systems, safe spaces to learn from failures, and opportunities to develop optimized problem-solving strategies. Customizable content and formative feedback based on user performance contribute to individualized learning experiences, highlighting the immense potential of games in education (Ishak et al., 2023). To further capitalize on this potential, researchers employ AI and machine learning technologies to enhance the design and development of both games and their corresponding instructional materials (Sunarya, 2022).

Although DGBL has proven effective in education, it requires further development to maintain participants' interests and motivations while providing timely assistance without disrupting immersion. One promising approach is incorporating intelligent virtual agents into the game, offering support based on users' needs (Tumenayu et al., 2014). Another critical aspect requiring AI and machine learning techniques is adaptive learning games, which emphasize the impact of personalization and personification elements within game design on learning enhancement (Dyulicheva et al., 2020). Researchers (Serhan et al., 2019; Mulwa et al., 2010; Zhu and Ontañón, 2020; Khenissi et al., 2013; Soflano et al., 2015) have explored the implementation of dynamically adaptive DGBL environments that depend on participants' learning styles, performance, in-game behavior, personality traits, rating scores, and preferences. Notably, AI and machine learning techniques enable the development of real-time adaptive learning games that flexibly adjust game content and difficulty based on learners' ongoing data during gameplay

(López and Tucker, 2018; Hooshyar et al., 2021). In DGBL with virtual and augmented reality (VR/AR) technologies, AI and machine learning techniques aid researchers and practitioners in developing adaptive hints and training materials (Drey et al., 2020) within the virtual environment to improve learning for learners with special needs (Dyulicheva et al., 2020) and in domains requiring a better understanding of abstract or complex ideas, contextual awareness, and problem-solving skills in high-stakes or hazardous situations (Dyulicheva et al., 2020; Afyouni et al., 2020; Lin et al., 2021).

Another area of DGBL involves students learning while developing games from scratch, transforming the role of instructors from passive observers to active participants (Kuznetsov et al., 2020). With the support of AI and machine learning tools, instructors can actively and appropriately participate in students' learning by directly interacting with students within the game world based on specific learning objectives. They can also interact with the virtual environment to create in-game objects and define how to interact with them (Carbonaro et al., 2006). AI and machine learning tools facilitate the incorporation of adaptive elements into game mechanics to enhance game dynamics and create individualized learning and play paths for users (Spronck et al., 2006). Additionally, during the game development process, participants can implement AI blocks allowing recognition of speech, images, videos, and handwritten text for better learning experiences and outcomes in fields such as language learning, art recognition, and arithmetic skills (`https://github.com/ecraft2learn/ai`, 2021). This learning-while-doing teaching method in DGBL is especially beneficial for teaching difficult-to-learn concepts (Estevez et al., 2019)), fostering collaborative skills (Annetta et al., 2006), developing creativity, and enhancing self-education (Carbonaro et al., 2006).

In summary, AI and machine learning applications within DGBL rely on precise performance assessment, either directly or indirectly, to develop adaptive learning experiences customized to each individual's characteristics and gameplay history. For example, intelligent virtual agents utilize assessment data to establish appropriate interactions with students, creating tailored guidance and materials that cater to their diverse abilities. Likewise, adaptive learning environments depend on assessment data to generate formative feedback, personalized scenarios, tasks, and content based on participants' in-game behaviors and scores. In DGBL contexts, instructors actively engage in students' learning processes, employing assessment scores to devise supplementary activities and interventions that foster individualized experiences. Performance assessment remains integral to shaping effective educational strategies and enhancing continuous skill development. However, accurately assessing students' in-game performances is an ongoing, complex challenge due to the dynamic nature of the game world and the abundance of noisy behavior records generated during gameplay.

2.2. Assessment within Game-Based Learning

Assessment practices in education generally present significant challenges, which primarily stem from the complexities associated with validating the constructs of the measured knowledge, skills, or competencies; accurately defining the ultimate purpose of the assessment, such as formative feedback, summative review, or program evaluation; achieving alignment between students' assessment responses and predefined measurement objectives; and applying a single assessment to students with diverse backgrounds and characteristics (Kim and Ifenthaler, 2019). In game-based assessments, these challenges are further intensified by stakeholders' additional expectations and requirements concerning DGBL platforms.

Advocates of DGBL argue that such environments should foster sustained learning by immersing users in interactive game worlds featuring extensive game mechanics. These environments guide learners through scaffolded tasks embedded with pedagogical activities, ultimately supporting the acquisition of new knowledge, skills, and competencies - particularly in complex and dynamic domains characterized by ill-structured problems (Gee, 2003; Ifenthaler et al., 2012; Prensky, 2001; Shaffer, 2006; Saleh et al., 2019). In contrast to traditional educational courses that necessitate memorization of abstract concepts and procedures without contextual understanding, DGBL platforms emphasize active learning by prompting learners to autonomously discover pertinent clues and materials dispersed throughout the game world and to apply the knowledge acquired from these resources in problem-solving situations.

Owing to these differences in instructional approaches, digital game-based assessment (DGBA) demands a more comprehensive range of measurement dimensions than their traditional counterparts. Examples of such dimensions include game skill, problem-solving ability, information retrieval and synthesis, and learning capacity. Additionally, stakeholders are keenly interested in examining students' learning progressions as they advance through the game and in exploring the influence of individual characteristics on learning during gameplay, which imbues DGBA with real-time dynamism and further amplifies its complexity.

Over recent decades, the field of assessment within DGBL has experienced significant growth, as numerous research studies have demonstrated the efficacy of video games as instructional tools. These games have been shown to enhance overall learning outcomes, encompassing cognitive and interpersonal skills (Clark et al., 2016; Boyle et al., 2016), as well as domain-specific knowledge, particularly in science and mathematics (Divjak and Tomić, 2011), when compared to conventional educational platforms. Throughout this period, researchers have predominantly employed external assessment methods, such as questionnaires, interviews, and observational records, to evaluate students' learning outcomes, engagement levels, and the game's performance, as well as its usability. In a systematic review conducted by Gris and Bengtson (2021) which focused on DGBA, the authors discovered that a mere 2.75% of studies measuring learning, 0.88% of studies measuring engagement, and 1.61% of studies measuring usability relied on data obtained from within the game, such as game logs. In their discussion, Gris and Bengtson emphasize the pressing necessity for future empirical research to construct validated and reliable assessments that draw upon data from within the game.

Since its inception in 2009, GlassLab (`http://www.instituteofplay.org/work/projects/glasslab-research/`) (Rowe et al., 2015) has garnered attention from scholars and practitioners due to the detailed data traces collected from its games. These data traces offer numerous events per learning activity, providing a remarkable opportunity to extensively analyze learning in diverse aspects. This analysis can potentially elevate assessment technologies, bolstering DGBL through data-driven methods that can be scaled to encompass the entire educational domain. This belief has sparked many research endeavors exploring DGBA in recent decades.

For example, Eseryel et al. (2011) proposed an embedded framework for assessing complex problem-solving in a longitudinal design-based research study. This framework relied on two methodologies, adapted protocol analysis and HIMATT, to generate quantitative measures and visualizations for instantaneous feedback during gameplay (Ifenthaler, 2014). Subsequently, researchers integrated techniques from learning analytics and educational data mining into DGBL analytics, advancing the research field with theoretical support and validated methods for interpreting results (Loh et al., 2015). Rowe et al. (2017) developed detectors using a machine-learned algorithm based on in-game log data to gauge implicit understanding of physics, identify strategies associated with in-game productivity, and assess computational thinking, enabling real-time player inferences. Similarly, Kim and Rosenheck (2020) employed sophisticated learning analytics and educational data mining techniques to guide the design and development of games for assessment purposes. Additionally, Tadayon and Pottie (2020) investigated the application of hidden Markov models on sequences of learning actions within a DGBL platform, confirming the efficacy of this approach for predicting learning outcomes.

2.2.1. Stealth Assessment within Game-Based Learning

Stealth assessment (SA) is a relatively mature and structured research line within the realm of DGBA. SA within DGBL environments has been the subject of continuous investigation in recent years, given its promising potential to meet the high expectations of stakeholders regarding DGBL. This research line is also capable of measuring learning progressions related to complex competencies, such as those identified as 21st-century competencies (Romero et al., 2015). These competencies are essential for enabling new generations to adapt to a rapidly changing world, and they are difficult to quantify through conventional assessment methods (e.g., paper-based exams, pre-post tests) or traditional educational platforms.

SA within DGBL platforms aims to unobtrusively evaluate participants' diverse performance metrics using extensive trace data gathered from adaptive logging systems embedded in the game. This approach preserves engagement and learning flow, as noted by Shute et al. (2009). SA is designed to deliver continuous, multifaceted information on learners inconspicuously, rendering the measurement process more objective and comprehensive. By harnessing advanced machine learning and artificial intelligence techniques, SA can perform real-time scoring based on students' actions and learning progress, providing accurate formative feedback. Moreover, through its sophisticated integration into game-based environments, SA measures learning in a context-aware manner, contributing to the advancement of adaptive learning within the realm of DGBL. Subsequent paragraphs present a concise, systematic review of SA within DGBL contexts.

Göbel and colleagues (Göbel et al., 2009; Göbel and Mehm, 2013) conducted preliminary long-term research on story-based edutainment applications and serious games, resulting in the development of a prototype framework for SA in story-based digital educational games (DEG) called Narrative Game-based Learning Objects (NGLOB). This framework was demonstrated and validated using two existing computer-based games; however, its applicability has waned in recent years, potentially due to restrictions in suitable game genres.

Shute (2011) developed a versatile SA model grounded in evidence-centered design (ECD) (Mislevy et al., 2003) and applied it to various DGBL environments. Their investigations covered a range of educational games, evaluating competencies such as mathematical abilities (Shute et al., 2017), problem-solving capabilities (Shute et al., 2016), conscientiousness (Moore and Shute, 2017), calculus proficiency (Smith et al., 2019), and creativity (Shute and Rahimi, 2021). The ECD SA model consists of three primary elements: the competency model, the evidence model, and the task model. These components enable practitioners to examine learning behavior patterns and estimate competence levels promptly. Shute's studies focus on discerning relationships between different in-game behavior-derived indicators and assessed competencies using Bayesian networks (BN). BNs effectively visualize complex relationships, including time factors, to keep data valuable and manageable (Mouri et al., 2016; Belland et al., 2017;

Champion and Elkan, 2017; Heine, 2020). However, developing BNs is labor-intensive, time-consuming, and costly (Belland et al., 2017), to ensure accurate representation of learning in the final structure. Furthermore, a tailored BN structure is often specific to a particular game environment, making it challenging to apply it to other contexts directly (Georgiadis et al., 2019).

Exploring beyond BNs, researchers working with Lester have examined the use of advanced machine learning models such as Random Forest, Support Vector Machine, and Recurrent Neural Networks for SAs within DGBL environments (Akram et al., 2018; Min et al., 2019; Henderson et al., 2020; Gupta et al., 2021; Henderson et al., 2022). They identified in-game behaviors linked to targeted knowledge and skills, integrating these models into the ECD framework. This resulted in novel SA frameworks with various benefits: streamlining data preprocessing (Min et al., 2019), enabling the operation of SAs in domains and educational content where prior data and labels are unavailable (Henderson et al., 2022), and infusing diverse data types (Henderson et al., 2020). However, challenges remain, including the limited interpretability of model outputs - such as understanding how individual indicators predict learning outcomes or identify game-based behaviors - and the difficulty of generalizing these approaches to dynamic, multi-faceted game environments (Akram et al., 2018; Min et al., 2019). Furthermore, while promising, these techniques often require additional refinement to balance accuracy, scalability, and usability in practical applications.

In summary, the rapidly evolving field of SA within DGBL environments has shown significant potential in evaluating a broad range of competencies. However, much of the research has focused on specific aspects of the DGBL environment to address scalability challenges or has constrained student interaction formats to simplify the complexities of measuring dynamic and multifaceted game worlds. While studies based on complete games exist, the generalizability of these methods to other DGBL contexts is often limited due to the reliance on manually crafted features or predictors tailored to specific environments (Shute and Rahimi, 2021)

Researchers have investigated the potential of automatically generating predictors from raw game logs using advanced machine learning models to mitigate the limitations of expert-engineered features—such as their complexity, labor-intensive nature, and time consumption. Although these efforts have produced several promising results, significant challenges persist. These include the loss of model interpretability (Min et al., 2019), a lack of empirical analysis in addressing ill-defined problems involving extensive in-game action formats (Akram et al., 2018), and generally lower accuracy rates than models utilizing expert-generated features. Botelho et al. (2019) demonstrated that models based on expert features consistently outperformed those relying on machine-learned features extracted from raw logs. Despite the absence of a universally accepted accuracy benchmark for machine learning models in SA within DGBL environments, it is evident that higher accuracy rates enhance the credibility of these models in practical applications. Furthermore, Botelho and colleagues introduced an innovative feature engineering method that integrates expert and machine-learned features, resulting in superior model performance compared to traditional techniques. Therefore, selecting a feature engineering method should be guided by the research objectives, weighing the trade-offs between model performance, interpretability, time, cost, and labor.

3. Methodology

As we reviewed in the previous section, several effective SAs grounded in the ECD approach have been implemented within various DGBL contexts (Shute et al., 2016; Min et al., 2019;

Smith et al., 2019; Shute and Rahimi, 2021; Henderson et al., 2022). However, the high-level nature of ECD, which serves primarily as a conceptual model for defining the broad components of competence, task, and evidence models, necessitates additional guidance for practitioners to organize and standardize elements within each component systematically. Many studies have relied on expert intuition for these decisions (Shute et al., 2016; Smith et al., 2019; Shute and Rahimi, 2021), underscoring the need for a more structured framework applicable to diverse DGBL environments. To ensure that SAs effectively and robustly measure the intended learning objectives, it is crucial to provide sophisticated and systematic guidance for defining the elements within all three ECD components.

Given the scope of this work, the focus will be on the evidence model, which outlines the generation of evidence (features) from game content and the statistical models (e.g., machine learning models) that link this evidence to learning objectives. The following sections present a pipeline designed to guide the analytical process, particularly within the Evidence Model, validated through empirical data collected from a DGBL environment—Mission HydroSci (MHS).

3.1. Digital Game-Based Learning Environment: Mission HydroSci

In this study, we utilized the DGBL tool Mission HydroSci (MHS), a 3D narrative adventure game designed for middle school students, to evaluate the effectiveness of our proposed analytical pipeline. MHS aligns with the NGSS, a framework that integrates three key dimensions—disciplinary core ideas, science and engineering practices, and crosscutting concepts—to provide students with a comprehensive and application-focused understanding of science. These standards aim to prepare students for success in college, careers, and civic life (National Research Council, 2013).

MHS is designed to teach water science and scientific argumentation through a transformative play approach (Barab et al., 2010), where players adopt the role of a character and apply in-game knowledge to solve real-world problems. The game's design incorporates a learning progressions methodology grounded in extensive research on water systems (Covitt et al., 2009; Sadler et al., 2017) and scientific argumentation (Osborne et al., 2013). Our assessment strategy leverages a sophisticated logging system, informed by the Activity Theory-based Model of Serious Games (ATMSG) (Carvalho et al., 2015) and the Experience API (xAPI) standards (Serrano-Laguna et al., 2017), to track and analyze students' in-game interactions. This system enables the development of adaptive assessment tools and the provision of tailored formative feedback, laying the groundwork for an adaptive learning system.

MHS consists of six modules, each focused on specific curriculum topics and featuring distinct virtual landscapes, game mechanics, and embedded educational materials. The game requires approximately 10 hours to complete within a classroom setting under the guidance of an instructor. Please refer to Appendix A for more detailed illustrations and descriptions of MHS.

3.2. Data Collection and Summary

3.2.1. Research Design and Data Collection

The data utilized in this study were collected during the second field test of MHS, conducted between February 11 and April 15, 2019. Prior to the test, thirteen middle school science teachers from nine schools across six school districts were recruited through notifications sent to science coordinators and the state science teachers association. All participating schools and teachers were located in a single Midwestern state, representing a mix of public schools from both mid-sized cities and small rural communities.

The student sample for this study (N = 806) was composed of 51% male and 49% female students, with a demographic breakdown of 66% Caucasian, 11% African American, 6% Hispanic, 4% identifying as multi-racial, 3% Asian, 2% American Indian, and the remaining students self-identifying as other. The study spanned ten school days, with the first and last days allocated for pre- and post-testing. These assessments measured students' knowledge of water systems, argumentation skills, and how playing MHS affected their attitudes toward learning science and technology. However, this study focuses explicitly on the pre- and post-assessment outcomes related to water systems knowledge and scientific argumentation.

All testing, including pre-and post-tests, was completed within a single class period, approximately 40 to 45 minutes. Students took the assessments online, with the science affect measure administered last to maximize time for the water systems and argumentation assessments. Of the 806 students, 632 completed the pre-and post-tests for all constructs and were included in the final analytic sample. Further details on the study methodology can be found in (Reeves et al., 2020).

3.2.2. Game Log Collection and Summary

In terms of game log collection, all 632 students' game logs should be collected in the MHS logging system. However, due to technical issues, such as internet loss, not all students' playthrough records are saved on the distant server. Moreover, although we provided assistance materials such as a dashboard, strategy guide, slides, and discussion questions to help teachers' lesson preparation, it is still comparatively new for them to incorporate a game into their pedagogical arrangements so that unexpected situations happen during courses. Consequently, the actual courses always fell behind the pedagogical arrangements, decreasing the completion rate for each unit. So, the sample numbers of training the machine-learning models for each unit differ. We only involved students' log records who passed all primary quests for a specific unit and have completed pre- and post-assessment records. Regarding students who may replay the game several times, we only included their first trials that complete all primary quests for a particular unit.

Table 1 lists detailed information regarding the number of students and corresponding log records involved in the model training.

Table 1: Log records information at scales of each unit and the whole game.

	Student Count	Sum log record	Average log record per student per trial
Unit 2	350	4,277,771	4,681
Unit 3	323	11,459,160	14,844
Unit 4	181	7,208,357	11,721
Unit 5	128	4,234,646	9,107
Whole game	463	34,320,924	31,574

3.2.3. The Measurement of Learning Outcomes: Pre- and Post-Assessment Test Results Summary

As mentioned, pre- and post-assessment tests administered before and after the gaming experience evaluated students' learning. Two external assessments were incorporated to gauge diverse aspects of learning: the Water Systems Assessment (WSA) for assessing content knowledge acquisition and the Argumentation Assessment (AA) for appraising scientific argumentation competencies. Both assessments comprised a series of multiple-choice questions. To investigate whether there was a significant score gain between the pre-and post-assessments, we conducted statistical tests and observed significant score enhancements in units 2, 3, 4, and 5 and also the aggregate scores for all items in both WSA and AA within the scope of the whole game. Units 1 and 6 were excluded from this analysis: Unit 1 is a tutorial unit designed to teach students how to play the game and introduce the story background, without specific curriculum content; Unit 6 is a summarization unit that was not fully developed at the time of the field test. More detailed information regarding the testing process and results can be found in Appendix B.

3.3. The Analytic Pipeline Description and Application Using Mission HydroSci

To address the limitations discussed earlier in the Methodology section and support the broader implementation of SAs within DGBL environments—particularly those involving complex and dynamic game worlds—we have developed a pipeline to guide the analytical process, specifically focusing on the evidence model. This pipeline is illustrated in Figure 1.

In the following subsections, we describe the specific elements within the pipeline and the analytical process applied to MHS following the outlined wpipeline. The objective is to predict students' learning outcomes in water science content knowledge across different units (curriculum topics) and their scientific argumentation skills after engaging with MHS. Each subsection corresponds to a specific element in the pipeline.

For the implementation process, we initially utilized R, specifically the "mongolite"[1] and "Tidyverse"[2] libraries. The "mongolite" library was employed to connect to MongoDB and retrieve raw logs from the server. At the same time "Tidyverse" was used for processing and manipulating these raw logs to generate the original features, as discussed in Section 3.3.5. Following this, we transitioned to Python, using Jupyter Notebook, and employed the "scikit-learn,"[3] "NumPy,"[4] and "Pandas"[5] libraries to complete the subsequent steps. "NumPy" and "Pandas" facilitated the necessary data-wrangling processes. At the same time "scikit-learn" was used to generate transformed feature sets and to conduct the machine learning model training, validation, and testing processes.

3.3.1. Element (1): Extracting Raw Logs from the Integrated Game Logging System

In this step, practitioners extract participants' game logs from the adaptive logging system integrated within the digital game, with the specific game content (e.g., game tasks or quests) determined by the ECD task model.

[1] mongolite: https://cran.r-project.org/web/packages/mongolite/index.html

[2] Tidyverse: https://www.tidyverse.org/

[3] scikit-learn: https://scikit-learn.org/stable/

[4] NumPy: https://numpy.org/

[5] Pandas: https://pandas.pydata.org/

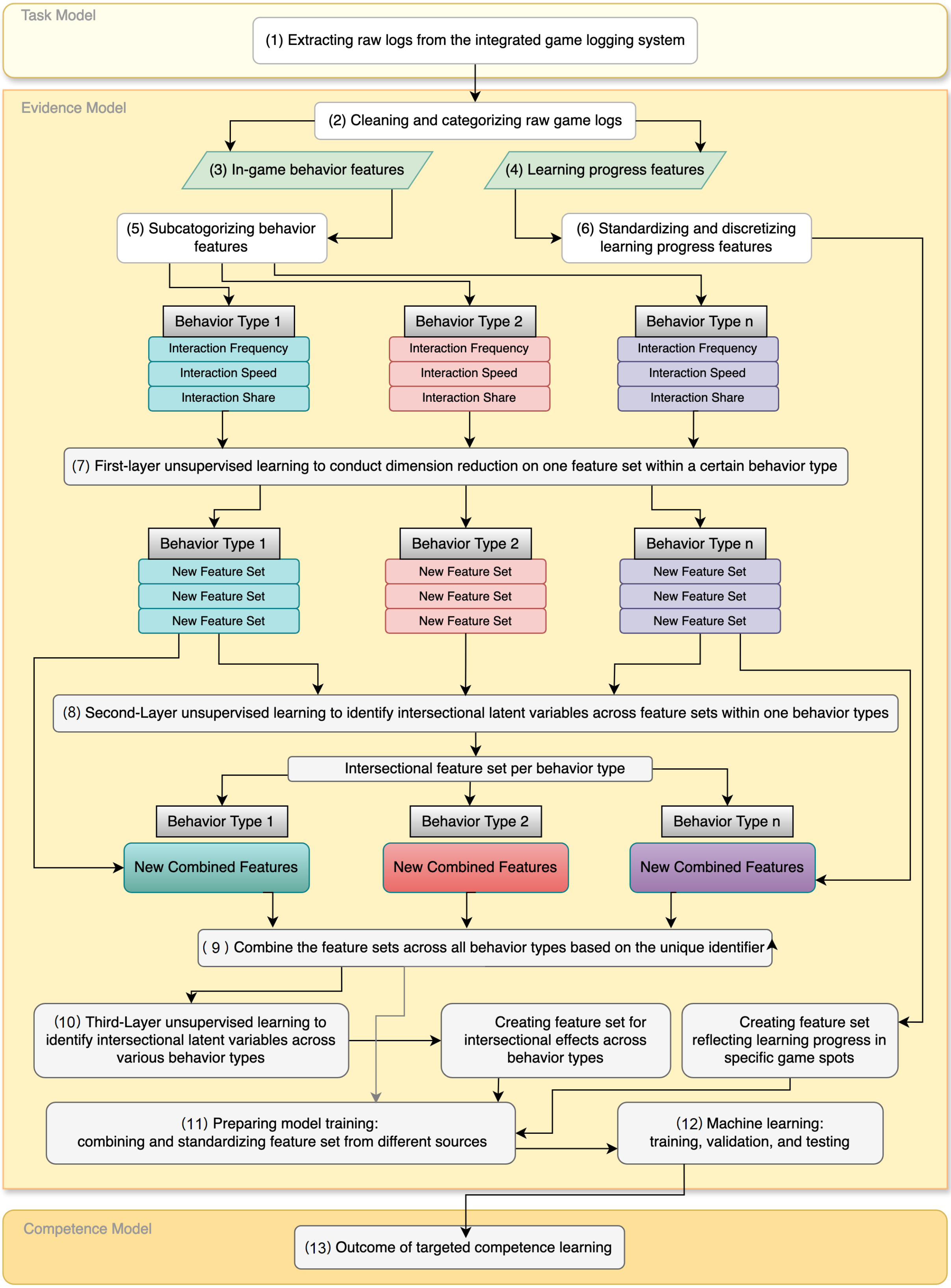

Task Model
(1) Extracting raw logs from the integrated game logging system
Evidence Model
(2) Cleaning and categorizing raw game logs
(3) In-game behavior features
(4) Learning progress features
(5) Subcatogorizing behavior features
(6) Standardizing and discretizing learning progress features
Behavior Type 1
Interaction Frequency
Interaction Speed
Interaction Share
Behavior Type 2
Interaction Frequency
Interaction Speed
Interaction Share
Behavior Type n
Interaction Frequency
Interaction Speed
Interaction Share
(7) First-layer unsupervised learning to conduct dimension reduction on one feature set within a certain behavior type
Behavior Type 1
New Feature Set
New Feature Set
New Feature Set
Behavior Type 2
New Feature Set
New Feature Set
New Feature Set
Behavior Type n
New Feature Set
New Feature Set
New Feature Set
(8) Second-Layer unsupervised learning to identify intersectional latent variables across feature sets within one behavior types
Intersectional feature set per behavior type
Behavior Type 1
New Combined Features
Behavior Type 2
New Combined Features
Behavior Type n
New Combined Features
(9) Combine the feature sets across all behavior types based on the unique identifier
(10) Third-Layer unsupervised learning to identify intersectional latent variables across various behavior types
Creating feature set for intersectional effects across behavior types
Creating feature set reflecting learning progress in specific game spots
(11) Preparing model training: combining and standardizing feature set from different sources
(12) Machine learning: training, validation, and testing
Competence Model
(13) Outcome of targeted competence learning

Figure 1: The diagram to outline the elements of the general analytical pipeline.

In the case of MHS, the integrated logging system operates on two layers. The first layer contains general information common to each log record. The second layer includes additional detailed information specific to each log record, which varies depending on the type of event or behavior. The variable "ItemID" links these two layers, and a unique identifier is assigned to each log record. For more comprehensive illustrations regarding our embedded logging system, refer to Appendix D.1 and our previous publication (Lu et al., 2023).

3.3.2. Element (2): Cleaning and Categorizing Raw Game Logs

After extraction, the raw game logs are processed and cleaned, resulting in data files structured into multiple rows and columns. Each row corresponds to a unique log record, and each column represents a specific record attribute. The game logs are then categorized into two data groups: participants' in-game behavior features (Element 3) and learning progress features (Element 4).

In applying this step to MHS, we removed students' log records that were not continuous, lacked sufficient data (e.g., cases where the game froze before completing the first main quest for each unit), or were generated when students replayed the same game content multiple times (only the log record from the first playthrough was retained). Based on the cleaned log dataset, we manually crafted features representing students' learning outcomes at specific points within the game, referred to as embedded assessment scores, which served as the learning progress features (discussed in Subsection 3.3.4). Other features were categorized as behavioral features (discussed in Subsection 3.3.3).

3.3.3. Element (3): In-Game Behavior Features

These features capture participants' rich and varied in-game actions, including their interactions with in-game objects, navigation within the game world, and responses to non-player characters (NPCs). For MHS, we generated feature sets that encompass various in-game behaviors such as task completion, argumentation, and map exploration.

3.3.4. Element (4): Learning Progress Features

These features are expertly generated to describe participants' learning outcomes in specific game segments or tasks. Typically, these expert-crafted features act as special markers, indicating how well a participant performs in a pedagogical activity within the game and whether the activity effectively imparts the intended knowledge or skill.

For MHS, the research team conducted a series of discussions to identify the game segments suitable as measurement points for assessing students' learning outcomes. Additionally, the team established corresponding standards, which were used to create a scoring rubric table. Due to space limitations, Table 2 presents a portion of this scoring rubric table used to generate the learning progress features. The complete scoring rubric table is available in Appendix C.

This study included 16 learning progress features in total: 4 for Unit 2, 3 for Unit 3, 4 for Unit 4, and 5 for Unit 5.

3.3.5. Element (5): Subcategorizing Behavior Features

Practitioners can further subcategorize behavior features based on specific actions, such as task completion, dialogue reading, and tool application. For MHS, we categorized behavior features into two segments based on their nature:

Table 2: Part of the scoring rubric table for generating learning progress features.

Unit	Corresponding Quest	Learning Progress Name	Calculation Standards

Unit 2	**Argue which watershed is bigger:** In this quest, students will enter into a 2-D system where they will generate a complete argumentation with three components - Evidence, Reasoning, and Claim – by dragging and dropping available choices.	biggerArgScore	**2 Points:** correct answer within 3 attempts;
			1 Point: correct answer within 4 attempts;
			0 Points: no correct answer or correct answer after more than 4 attempts.

	CREI system: In this quest, students will enter a new game area where they are asked to deliver or kick a soccer ball into different directions. Each direction represents a component of a complete argumentation. Students need to make the right decision based on the information they got from dialogues with an in-game NPC.	CREIScore	**1 Point:** for each correct soccer ball delivery, students will get one score for this quest; **-1/3 Points:** for each incorrect soccer ball delivery, students will lose 1/3 point.

ACTION-SPECIFIC BEHAVIORS These behaviors represent direct interactions tied to specific game mechanics or tasks: (1) **Task completion** behaviors focus on two key metrics: the time spent completing individual tasks (interaction speed) and the proportion of total gameplay time allocated to each task (interaction share). (2) **Argumentation** behaviors are analyzed by measuring the frequency of student interactions within the argumentation system—such as dragging, dropping, hovering over elements, and using tools—along with the average time spent on these actions. These measures help assess both the frequency and efficiency of students' engagement with argumentation tasks. (3) **Hotkey usage** captures the frequency with which students utilize hotkeys for rapid access to game functions, providing insights into how efficiently they navigate the game. (4) **Tool menu usage** reflects students' reliance on the AI assistant, ARF, with metrics capturing both the frequency of tool usage and the time spent referring to each tool. (5) **Dialogue reading** is evaluated by tracking the frequency and speed at which students read in-game dialogues. (6) **Item triggering** measures the frequency of interactions with various in-game items and the duration of these interactions, particularly for items that involve prolonged engagement, such as those involved in puzzle-solving.

OVERARCHING BEHAVIOR CATEGORIES These behaviors represent broader patterns or meta-level actions across multiple game mechanics: (1) **Behavior type statement** captures overarching categories of behavior—such as movement, tool usage, or dialogue reading—and quantifies how frequently each type occurs and the time students spend on them. (2) **Map exploration** assesses the extent to which students explore the game world by measuring both the percentage of the map they uncover (interaction share) and the time they spend in different game areas (interaction speed).

Within each behavior type, feature sets can be generated to quantify the behavior from three perspectives: (1) Interaction frequency, which records how often a specific behavior occurs; (2) Interaction speed, which measures the average time spent on each occurrence of the behavior; and (3) Interaction share, which assesses the proportion of a specific behavior relative to the total frequency of all possible behaviors within that category. Importantly, not all three feature sets need to be included for every behavior type. Their inclusion depends on the characteristics of the feature sets (e.g., sufficient variance) and the relevance of the game content (e.g., if the speed of opening a door is consistent and needs more research significance, it may not be included).

The final feature set for each behavior type should be as comprehensive as possible to ensure sufficient information for subsequent model training. Collectively, these behavioral measures provide a detailed view of how students interact with the game's mechanics and features, offering insights into their learning processes and gameplay strategies. More detailed descriptions of these behaviors can be found within Table 10 of Appendix D.2.

Since the behavior types listed above are highly tailored to the specific context of the MHS case, we cannot assert that this list is exhaustive for all future studies or game environments. Other DGBLs may likely require customization of behavior features, either by adjusting the level of granularity or by introducing new behavior types unique to their context. We recommend that future practitioners use the behavior types outlined here as a starting point for feature generation in the early stages of their study. Based on their own findings, they can then customize the feature set in subsequent stages, adding or refining behavior types to better align with the specific learning objectives or mechanics of their game.

3.3.6. Element (6): Standardizing and Discretizing Learning Progress Features

For the learning progress features generated in Element 4, practitioners may need to preprocess these features using techniques such as scaling, normalization, or discretization, depending on the statistical characteristics of each feature. In the case of MHS, we applied discretization to the learning progress features, specifically the embedded assessment scores. Each score was categorized as either "High-Score" or "Low-Score" based on its comparison to the average score. If a student's score exceeded the average level, it was classified as "High-Score"; otherwise, it was categorized as "Low-Score." The final output of this step is a standardized feature set that reflects participants' learning progress across different game spots. This feature set can be integrated with in-game behavior feature sets in subsequent steps of the pipeline.

3.3.7. Element (7): First-layer Unsupervised Learning to Conduct Dimension Reduction on One Feature Set Within a Certain Behavior Type

Dimensionality reduction, a key function of unsupervised learning, is essential for addressing challenges when working with high-dimensional datasets. This process involves extracting features directly relevant to machine learning tasks, facilitating knowledge discovery and pattern classification among numerous redundant or irrelevant features. Additionally, dimensionality reduction techniques help reduce high-dimensional datasets to lower-dimensional representations by filtering out or removing redundant and noisy information, which can significantly improve the performance of computational models (Zebari et al., 2020). These techniques are generally categorized into two main groups: feature selection and feature extraction. Since feature selection may lead to considerable information loss by eliminating features that do not significantly contribute to predictions, we believe feature extraction techniques are more appropriate for analyzing how players' game logs influence their learning outcomes. Feature extraction reduces dimensionality with minimal information loss from the original dataset (Ayesha et al., 2020; Abd-Alsabour, 2018; Verleysen and François, 2005; Huang et al., 2019).

Due to the complex and dynamic nature of DGBL environments, a single feature set within a specific behavior type may contain hundreds or thousands of columns. These columns need to be carefully examined to extract relevant information that can be used to predict targeted learning outcomes. The first-layer unsupervised learning technique is applied to each feature set derived from Element 5 within a specific behavior type to extract relevant information and eliminate noise within the original feature set. After this process, each feature set will be transformed into a new one containing features generated by a specific unsupervised learning algorithm.

When selecting the most suitable unsupervised learning algorithm, practitioners can base their decision on the characteristics of the original feature set, such as scale, distribution, dimensionality, type of relationship (e.g., linear or nonlinear), presence of outliers, and the size of the feature set. Alternatively, practitioners may set up experiments to select the best algorithm from a pool of candidates.

To set up such experiments, practitioners could apply each unsupervised learning candidate to the same original feature set, generating transformed feature sets. These transformed sets are then evaluated using the same supervised learning model, such as logistic regression, to predict the targeted learning outcomes. The supervised learning model training should follow a 10-fold cross-validation process. Model performance can be assessed using the mean values of selected metrics, such as accuracy, precision, recall, and F1-score. Based on these evaluation metrics, practitioners can determine the best unsupervised learning algorithm to apply to the original feature set within a specific behavior type and decide how many transformed features should be included in the final feature set.

For MHS, we employed feature extraction techniques, a subset of unsupervised machine learning, on each feature set within a specific behavior type to reduce dimensionality as a preprocessing step before machine learning model training. Given the unique characteristics of each processed feature set, we determined the best feature extraction technique through an experimental comparison for each feature set rather than identifying a single overarching best technique. The methods we considered included Principal Component Analysis (PCA) (Abdi and Williams, 2010), Singular Value Decomposition (SVD) (Klema and Laub, 1980), Independent Component Analysis (ICA) (Hyvärinen and Oja, 2000), Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999), Kernel PCA (Schölkopf et al., 1997), T-distributed Stochastic Neighbor Embedding (t-SNE) (Wattenberg et al., 2016), U-map (McGaghie and Harris, 2018), and Autoencoders (Wang et al., 2016). A brief rationale for selecting each technique is provided in Appendix D, Table 11.

To conduct the experiments, we first evaluated a logistic regression model on each raw feature set—interaction frequency, interaction speed, and interaction share—within a specific behavior type, predicting the targeted learning outcome in the relevant game context from which the raw feature set was derived, as a benchmark. The evaluation metric was the mean and standard deviation of classification accuracy across all folds and repeats after 10 stratified 10-fold cross-validations. We then applied each of the eight selected feature extraction techniques to the raw feature sets to obtain transformed feature sets, using the same evaluation method as for the raw feature sets. For that specific feature set, we selected the feature extraction technique that produced a mean accuracy higher than the benchmark and significantly outperformed the other methods. Finally, we merged all transformed feature sets for each behavior type using the unique identifier playerID.

It is important to note that the selection of feature extraction techniques was context-specific and tailored to the unique properties of the feature sets within MHS. Consequently, the optimal technique for a given feature set in this study may not generalize to other feature sets or game contexts. This highlights the importance of flexibility in applying feature extraction methods, as practitioners must consider the distinct characteristics of each feature set to identify the most appropriate technique.

3.3.8. Element (8): Second-layer Unsupervised Learning to Identify Intersectional Latent Variables Across Feature Sets Within One Behavior Type

Incorporating latent features representing intersectional effects across different feature sets into machine learning model training can provide several distinct benefits. These enhancements improve the performance and robustness of computational models, mainly when dealing with complex datasets where interactions between features are subtle but critical for accurate predictions. Specifically, the key advantages include:

1. **Capturing Hidden Relationships:** Latent features can reveal hidden relationships between various feature sets that may not be apparent in the original data, offering a more comprehensive view and richer insights.

2. **Improving Predictive Accuracy:** Models trained with original and intersectional latent features often achieve higher predictive accuracy by utilizing both direct and derived insights, making them more resilient to variations in the data.

3. **Enhancing Generalization:** Latent features frequently represent invariant aspects of the datasets—features consistent across different conditions or domains. This can enhance a model's generalizability when applied to unseen data or varying conditions.

The advantages of integrating latent feature sets representing intersectional effects across various feature sets or domains have been empirically supported in previous studies (Nickel et al., 2015; Calixto et al., 2019; Wu et al., 2019; Bauer et al., 2022; Wang et al., 2022). Following Element 7, practitioners first obtain new feature sets within each behavior type and then combine them using a unique identifier, such as playerID. After this combination, practitioners can apply a second layer of unsupervised learning to extract intersectional effects among features within a specific behavior type. The choice of the unsupervised learning algorithm for this step may differ from the one used previously, as the characteristics of the new combined feature set may change. Practitioners can select a different algorithm based on their research needs and the characteristics of the combined feature set, or they can follow the selection process described in Element 7. In the case of MHS, we applied factor analysis to derive more comprehensive insights from the combined transformed feature sets within a particular behavior type. This method was used on the combined transformed feature set obtained from the previous step (Element 7) to generate a new feature set representing the intersectional effects across feature sets within that behavior type. Factor analysis is particularly suited for identifying latent variables influencing observed variables or features. When applied to transformed features, it can uncover underlying structures or hidden factors that were not fully captured in the initial rounds of feature extraction (Bartholomew et al., 2011).

The final output of this step is a consolidated intersectional feature set per behavior type, which integrates the combined transformed feature sets from Element 7 and incorporates the intersectional latent variables. This comprehensive feature set can then be used for subsequent processing and machine learning model training.

3.3.9. Element (9): Generating a New Feature Set That Combines New Feature Sets of All Behavior Types

In this step, practitioners combine all the new feature sets generated from Element 7 with the feature set representing intersectional effects identified in Element 8. This combination is based on a unique identifier to form a comprehensive "Behavior Type x New Features" dataset. Practitioners combine the new features from all behavior types using the same unique identifier.

For MHS, we first merged the combined transformed feature set from Element 7 with the feature set from Element 8 to create a new feature set for each behavior type, using playerID as the key. Subsequently, we combined these feature sets across all behavior types using playerID to prepare the data for further analysis.

3.3.10. Element (10): Third-layer Unsupervised Learning to Identify Intersectional Latent Variables Across Various Behavior Types

With the combined feature set generated from Element 9, practitioners can apply unsupervised learning algorithms to identify intersectional latent variables across all behavior types. Given that both elements deal with latent variables, practitioners may either use the same algorithm selected in Element 8 or follow the experimental process described in Element 7 to choose an appropriate algorithm. For MHS, we once again employed factor analysis on the combined feature set from Element 9 to derive a new feature set representing the intersectional effects across all behavior types. The final output of this step is a feature set that encapsulates the intersectional effects across behavior types, providing a comprehensive representation for subsequent processing and machine learning model training.

3.3.11. Element (11): Preparing Model Training: Combining and Standardizing Feature set from Different Sources

The initial step in preparing for model training is to combine the feature sets from Elements 6, 9, and 10 using a unique identifier. Once combined, practitioners should preprocess the final dataset, including tasks such as standardization, imputation, and outlier exclusion. For MHS, we first merged the datasets from Elements 6, 9, and 10 using the unique identifier, playerID. Following this, we standardized the combined feature set to ensure that each feature was scaled consistently, which is crucial for effective model training.

3.3.12. Element (12): Machine learning: Training, Validation, and Testing

Based on the preprocessed dataset derived from Element 11, practitioners can proceed with training, validating, tuning, and testing machine learning models. For the MHS case, the process was as follows:

Given our feature set's emergent and context-dependent nature and the need for robust and generalized predictions across diverse data conditions, we decided to employ an Ensemble Learning approach using a hard voting scheme (Assiri et al., 2020). To optimize the hard-voting ensemble for binary classification, we selected classifiers that are diverse, robust, and complementary in their strengths. These classifiers included C-Support Vector Classification (SVC) (Cervantes et al., 2020), Random Forest (Breiman, 2001), Logistic Regression (Hosmer Jr et al., 2013), K-Nearest Neighbors (KNN) (Peterson, 2009), Gaussian Naïve Bayes (Kamel et al., 2019), XGBoost Classifier (Chen and Guestrin, 2016), Gradient Boosting Classifier (Natekin and Knoll, 2013), AdaBoost Classifier (Ying et al., 2013), Linear Discriminant Analysis (LDA) (Tharwat, 2016), and Quadratic Discriminant Analysis (QDA) (Tharwat, 2016). The rationale for selecting these classifiers is detailed in Appendix D, Table 12.

To ensure the stability, validity, and robustness of the model, we allocated 80% of the observations to the training dataset and the remaining 20% to the testing dataset, using 10 different random seeds (each seed represents a distinct way to split the dataset, ensuring variability in training and testing datasets).

Since the post-test assessment scores were discretized into "high" and "low" categories based on the average score, these two categories' sample sizes were imbalanced, potentially affecting model outcomes. To address this issue, we used stratified sampling when splitting the training and testing datasets to maintain the proportion of each class within both datasets. We also employed subsampling techniques on the training dataset, specifically using a hybrid approach, SMOTETomek (Sasada et al., 2020), which combines oversampling the minority class with SMOTE and Tomek Links to remove samples that contribute to class overlap.

During model training, we implemented 10-fold stratified cross-validation, which preserves the proportion of each class ("low" and "high") in each data split. We repeated this process 10 times to estimate model parameters, conduct feature selection, and assess performance.

We aimed to mitigate bias and overfitting for the feature selection process by selecting classifiers capable of providing feature importance scores from those involved in the ensemble learning model. These classifiers included Random Forest, XGBoost, Gradient Boosting, and AdaBoost. Each classifier was trained separately on the processed (standardized and encoded) training dataset. After training, we retrieved the feature importance scores from each model. We then normalized these scores across classifiers to ensure they were on the same scale and calculated the average importance score for each feature. Based on these average scores, we ranked the features.

We followed a forward feature selection process, beginning with a single feature (e.g., the pre-assessment score of a specific unit or the entire game) and adding additional features one by one, starting with the one with the highest average importance score. After adding each feature, we evaluated its contribution to the ensemble learning model's performance using the average F1 score from the cross-validations. If the F1 score improved, the feature was permanently added to the final feature set; if not, the feature was discarded, and we moved on to the next one.

After completing the training and validation process, we tested the trained model on the test dataset. Model performance was assessed using the average and standard deviation metrics such as accuracy, precision, recall, and F1 score across different dataset splits. Additionally, to compare model performance across the "Low" and "High" categories of learning outcomes (both per unit and across the entire game), we calculated these metrics for each class individually. The results section provides a detailed explanation of each performance metric.

3.3.13. Element (13): Outcome of Targeted Competence Learning

Within this element, practitioners should first clearly define the specific competencies, knowledge, or skills that their models are intended to predict, aligning with the learning goals outlined in the ECD competence model. Then, practitioners should appropriately measure and quantify these learning outcomes. For example, in our case, the dependent or target variable predicted by the classification model was generated by extracting relevant items from the post-test assessment. These items were summed to produce post-test scores aligned with the targeted learning goals (detailed information, including the number of items per unit and per competence within the assessment instruments, is provided in Table 8 of Appendix B). The summed scores were then discretized into "Low" and "High" categories based on the average score.

To gain insights beyond performance metrics, such as identifying key game behaviors that drive learning outcomes and adjusting game and pedagogy design accordingly, practitioners could consider conducting model inference to further interpret the model results. For MHS, we used the following approaches: 1) Model performance comparison: standard performance metrics on the test set allow practitioners to compare model performance across different game contexts and establish how well the model predicts learning outcomes. 2) Feature importance: Practitioners can calculate feature importance rates using permutation importance. This method identifies which in-game features contribute most to the model's predictions by evaluating how shuffling feature values impacts accuracy. 3) ALE plots: Accumulated local effects (ALE) plots can help visualize how individual features influence predictions, allowing practitioners to understand the nonlinear or interaction effects of in-game behaviors on learning outcomes.

3.4. Visualizing the Inference of Latent Variables from Observable Behaviors

This section illustrates the transformation of observable behaviors—captured through players' in-game actions—into latent variables using our structured, multi-layer analytic pipeline. It also demonstrates the relationship between observable features and the latent features generated at different stages.

For better illustration, we created Figure 2, reflecting the features generated from different pipeline stages and the relationships between those features. The following paragraphs contain comprehensive illustrations regarding the above figure.

3.4.1. Observable Variables

Observable variables are directly derived from players' interactions with the game environment, encompassing behaviors such as task completion, argumentation, hotkey usage, tool menu navigation, and in-game item interactions. These raw features are categorized into the types of Frequency, Speed and Share, as described in Section 3.3.5.

The observable features follow a standardized naming convention that reflects the behavior type, feature type, and sequence. For instance, raw interaction frequencies, speeds, and shares are expressed as:

- `(behavior type name)_frequency_raw_(feature name)_(number)`
- `(behavior type name)_speed_raw_(feature name)_(number)`
- `(behavior type name)_share_raw_(feature name)_(number).`

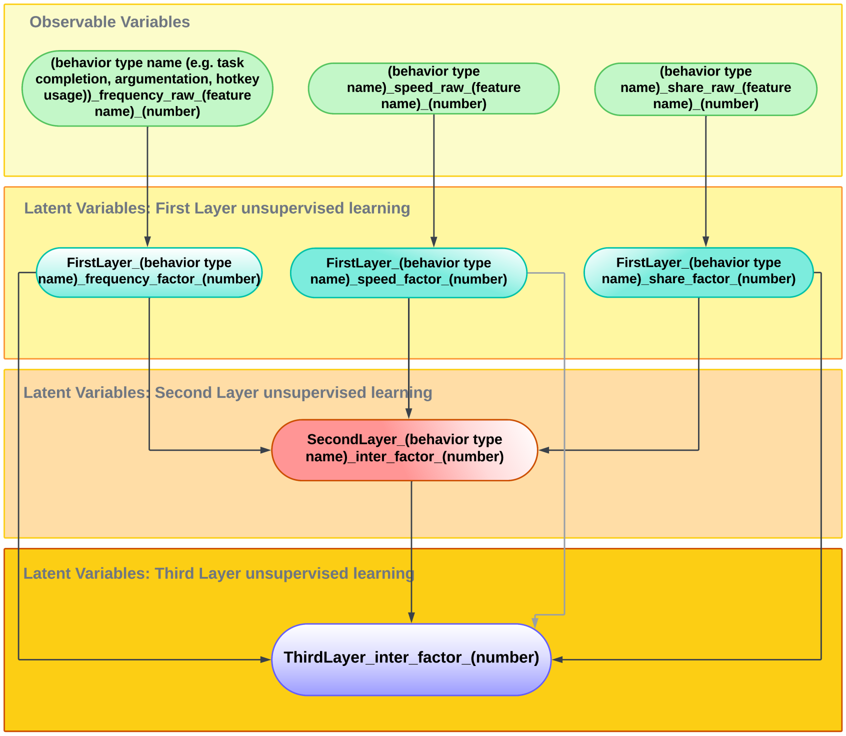

Observable Variables
(behavior type name (e.g. task completion, argumentation, hotkey usage))_frequency_raw_(feature name)_(number)
(behavior type name)_speed_raw_(feature name)_(number)
(behavior type name)_share_raw_(feature name)_(number)
Latent Variables: First Layer unsupervised learning
FirstLayer_(behavior type name)_frequency_factor_(number)
FirstLayer_(behavior type name)_speed_factor_(number)
FirstLayer_(behavior type name)_share_factor_(number)
Latent Variables: Second Layer unsupervised learning
SecondLayer_(behavior type name)_inter_factor_(number)
Latent Variables: Third Layer unsupervised learning
ThirdLayer_inter_factor_(number)

Figure 2: The diagram to outline the relationship between observable features and latent features.

These observable variables form the foundational data used for subsequent processing in the unsupervised learning pipeline.

3.4.2. Latent Variables: First Layer of Unsupervised Learning

The first layer of the unsupervised learning process (Element 7) applies dimension reduction techniques to the observable features, generating behavior-specific and feature-type-specific latent variables. This layer focuses on extracting key patterns from the raw data, creating a set of latent factors (i.e., frequency, speed, share) for each behavior type. The naming convention reflects these transformations:

- `FirstLayer_(behavior type name)_frequency_factor_(number)`
- `FirstLayer_(behavior type name)_speed_factor_(number)`
- `FirstLayer_(behavior type name)_share_factor_(number)`.

These first-layer latent variables capture condensed and abstracted representations of the raw data, enabling the identification of essential behavior patterns while maintaining the behavioral integrity of the observable features.

3.4.3. Latent Variables: Second Layer of Unsupervised Learning

The second layer of processing (Element 8) further abstracts the data by identifying intersectional latent variables across different feature types under a certain behavior type. This stage aggregates behavior-specific latent variables from the first layer, reflecting complex interactions between different feature types and uncovering intersectional effects. These intersectional latent variables are denoted as:

- `SecondLayer_(behavior type name)_inter_factor_(number).`

The second layer provides a deeper level of insight by examining how various behaviors interact and co-occur, offering a richer understanding of player behavior that transcends individual actions.

3.4.4. Latent Variables: Third Layer of Unsupervised Learning

In the third and final layer (Element 10), the model integrates second-layer latent variables to form high-level latent constructs that span multiple behavior types. These constructs represent the most condensed and abstracted set of features, encapsulating the comprehensive behavioral patterns exhibited by players across various in-game interactions. The naming convention for these latent constructs is as follows:

- `ThirdLayer_inter_factor_(number).`

At this stage, the latent variables provide a holistic view of the player's behavior, combining data from multiple interactions to infer overall competence and learning outcomes.

Finally, the dataset generated from the entire analytic pipeline—comprising latent features from the first, second, and third layers of unsupervised learning, along with processed features reflecting students' learning progress—serves as the input for machine learning model training.

3.5. Applying the MHS Case to the Development of an Evidence-Centered Design Framework

Given the unique curriculum topic of each MHS unit and our intent to empirically evaluate our proposed pipeline under diverse scopes or contexts within MHS, we established an ECD framework for each context. According to highly cited literature (e.g., Mislevy et al. (2003), Shute (2011), Shute et al. (2016)) illustrating the ECD framework and its corresponding applications in SA construction within DGBL environments, ECD can be generally categorized into task, evidence, and competency models. Specifically, the **task model** involves the tasks or situations in which evidence about the learner's competencies is elicited. It consists of the design of the environment and the activities that will generate the necessary data. The **evidence model** describes how practitioners can infer learners' competency from in-game performances or behaviors. It bridges the competency and task models by defining the observable variables and scoring rules. The **competency model** contains the knowledge, skills, and abilities the assessment intends to measure. By carefully defining content within each of the three components, practitioners can ensure the alignment between measurement contexts and objectives of the SA (Shute, 2011; Shute and Ventura, 2013; Shute and Kim, 2014; Shute et al., 2016; Min et al., 2019). Table 3 describes how we constructed the ECD framework under different contexts:

Table 3: Defining evidence-centered design (ECD) framework for each context.

Contexts	Task Model	Evidence Model	Competency Model
Unit 2	All main quests in unit 2 of MHS. (Detailed information about each quest in each unit can be found in Appendix A.)	**Observable variables:** 36 features are included after the feature selection process, of which 3 represent learning progress and 33 represent in-game actions.	The sum of posttest assessment scores related to the content knowledge taught within unit 2, which has 6 items in total.
Unit 3	All main quests in unit 3	**Observable variables:** 45 features are included after the feature selection process, of which 3 represent learning progress and 42 represent in-game actions.	The sum of posttest assessment scores related to the content knowledge taught within unit 3, which has 3 items in total.
Unit 4	All main quests in unit 4	**Observable variables:** 24 features are included after the feature selection process, of which 1 represents learning progress and 23 represent in-game actions.	The sum of posttest assessment scores related to the content knowledge taught within unit 4, which has 4 items in total.
Unit 5	All main quests in unit 5	**Observable variables:** 27 features are included after the feature selection process, of which no features represent learning progress within the final feature set.	The sum of posttest assessment scores related to the content knowledge taught within unit 5, which has 10 items in total.
Whole Game Content Knowledge	All main quests from unit 1 to unit 5	**Observable variables:** 45 features are included after the feature selection process, of which 3 represent learning progress and 42 represent in-game actions.	The sum of posttest assessment scores related to the content knowledge taught in the whole game, which has 23 items in total.
Argumentation Skill	All main quests from unit 1 to unit 5	**Observable variables:** 36 features are included after the feature selection process, of which 2 represent learning progress and 34 represent in-game actions.	The sum of posttest assessment scores related to the scientific argumentation taught in the whole game, which has 12 items in total.

As part of the evidence model within the ECD framework, we employed a hard-voting ensemble learning classifier as the statistical model. This method combines multiple classifiers to improve model robustness and predictive accuracy across all contexts. To avoid redundancy, this statistical model is applied consistently throughout the study and is not repeatedly defined in Table 3.

Notably, the final feature set selected after each round of model training, validating, and testing may be different because of the different thresholds of splitting the dataset (different random seed numbers). To optimize the model's robustness, stability, and generalizability and avoid the over-complexity of the model, the features listed in Table 3 are those that have consistently been selected as important predictors across different random seeds.

3.6. Summary

This section presents and applies an analytical pipeline designed to predict learning outcomes across multiple learning objectives within a complex DGBL environment. This pipeline introduces several unique and innovative approaches that distinguish it from previous methods:

Firstly, the pipeline integrates both in-game behavior features and learning progress features. This integration facilitates the extraction of rich, nuanced features that capture students' intricate in-game actions and decision-making processes and strategies. These features form a critical foundation for constructing accurate models to predict learning outcomes. Moreover, including expert-crafted learning progress features enhances the interpretability of these models, making them more meaningful and applicable within educational theory and practice.

Secondly, the pipeline systematically deconstructs complex in-game behaviors into distinct types, each representing a unique aspect of player interaction. This decomposition ensures that the multifaceted nature of player behaviors is comprehensively captured. These datasets are meticulously designed to highlight specific dimensions of behavior patterns by examining the occurrence of behaviors and their speed and relative importance.

Lastly, the pipeline employs a three-layered unsupervised learning approach. This multi-layered method facilitates dimension reduction and the identification of intersectional latent variables within and across different behavior types, eliminating noises within feature sets and adding significant depth to the analysis. This approach ensures that the most relevant features are identified and utilized effectively in predicting learning outcomes.

4. Results

4.1. Model Performance Evaluation

4.1.1. Comparison across Different Feature Sets

To assess the efficacy of our proposed pipeline in the empirical study utilizing MHS, we initially compared test accuracy rates across varying feature sets. The outcomes of this comparison are presented in Table 4.

Table 4: Comparison of mean testing accuracy rates across 10 different random seed numbers, used to split the training and testing datasets, among various contexts. The table includes five performance measurements illustrating the effectiveness of the proposed analytic pipeline. Detailed descriptions of these measurements can be found in the main text. The values in parentheses represent the standard deviation of the corresponding testing accuracy rates.

Game sections	Majority-based accuracy	Only pre-score accuracy	Original feature set accuracy	No intersectional features accuracy	Post-pipeline feature set accuracy
Unit 2	57%	64.62% (0.052)	67.2% (0.0685)	76.45% (0.028)	**84.3% (0.0377)**
Unit 3	69%	69.47% (0.074)	70% (0.0716)	78.22% (0.024)	**82.6% (0.0222)**
Unit 4	67%	67.88% (0.076)	67.57% (0.0985)	83.08% (0.02)	**87.4% (0.0347)**
Unit 5	73%	73.7% (0.076)	76% (0.1171)	86.56% (0.025)	**93.67% (0.0398)**
Whole Game	60%	60.22% (0.038)	73.72% (0.066)	81.81% (0.027)	**86.2% (0.029)**
Argumentation	55.5%	65.59% (0.08)	66.26% (0.065)	80.63% (0.023)	**85% (0.03)**

More specifically, **majority-based accuracy** measures how well a model performs concerning just predicting the majority class for every instance. If the model's testing accuracy is not significantly higher than the majority class predictor, it may not effectively leverage the features to make predictions. It is pertinent to note that the majority class in this study consistently represents the "high-level" class across all scenarios, which is to say the accuracy rate based on the majority class corresponds directly to the proportion of the "high-level" class. From Table 4, we can see that the accuracy rates under this column are lower than those of other columns.

For **only pre-score accuracy**, we only used the pre-assessment score as the predictor to establish another baseline model for comparison. By investigating the model performance based on this feature set, we can know whether the pre-assessment score alone is a strong predictor for the targeted variable, whether we need additional features as predictors to solve the classification task, and if the increases of students' learning outcomes are primarily because of they have better pre-knowledge regarding the teaching materials within MHS. The third column of Table 4 shows an accuracy increase from the majority-based rates across all scenarios. However, the degree of these increases varies considerably across the range of scenarios. As we can see, the water science knowledge accuracy rates in units 3, 4, and 5, and the whole game scope increased by less than 1%. In comparison, the accuracy rates of the water-science knowledge in unit 2 and the scientific argumentation skill increased by over 7%. Overall, none of the accuracy rates surpass 75%.

Regarding **original feature set accuracy**, compared to the third column, the accuracy rates utilizing the feature set merging combined feature sets described in Element 5, the feature set described in Element 4, and the pre-assessment score together do not universally indicate an increase across all conditions, especially for the water-science knowledge of unit 4, which even shows a decrease compared to the third column. We think it may be caused by the curse of dimensionality, which contains many redundant or irrelevant features. In scenarios indicating increased accuracy rates, the increase is relatively moderate within the water-science knowledge of units 2, 3, and 5 and scientific argumentation skills. A notable increase of over 10% is observed for the water-science knowledge in the whole game scope, compared to its counterpart of "only pre-score accuracy."

In terms of **no intersectional features accuracy**, which utilized the combined feature sets, each of which is processed after Element 7 merged with the feature set described in Element 6 but without including any intersectional features, compared to the fourth column, we can see a universal increase in accuracy rates across various scenarios with around 10%. Four out of six scenarios – the water science knowledge of unit 4, 5 and the whole game scope, and the scientific argumentation skill – have testing accuracy rates surpass 80%. Furthermore, the standard deviation of the testing accuracy rates significantly decreased, compared to the third and fourth columns, which we think is because dimension reduction techniques help to exclude redundant features and alleviate noise for better capture of useful patterns underlying the dataset.

For **post-pipeline feature set accuracy**, which used the final feature set from going through all 13 processes described in the pipeline, we can see a universal increase in the testing accuracy rates compared to the counterparts when using the feature set without intersectional features. All testing accuracies surpass the threshold of 80%, which is an acceptable prediction accuracy rate within educational contexts (Bird et al., 2021). However, we can also notice that, in 5 out of 6 scenarios, the standard deviations of this column are more significant than their counterparts of the fifth column. We think the possible reasons could be although involving intersectional features helps us capture useful information across different feature sets and all behavior types for better performance of our classification tasks, this kind of process could also introduce additional variabilities that increase the standard deviation of the testing accuracies across different dataset splitting thresholds (random seeds' numbers).

Regarding the comparison across different scenarios based on the "post-pipeline feature set accuracy," when focusing on water science content knowledge, we find that the highest mean test accuracy is seen in unit 5, whereas unit 3 is the lowest. The mean test accuracy for the whole game presents an intermediate performance level compared to the counterparts of other unit-based records. When comparing under the scope of the entire game, the mean test accuracy for the water science knowledge is higher than that of the scientific argumentation skill.

4.1.2. Comparison Using Multiple Measurement Metrics for Evaluating Model Performance

As previously discussed, the dependent variables, namely posttest assessment scores, measured students' learning outcomes in the current study and have been discretized into "high" and "low" levels based on their mean values. This methodology could introduce significant challenges related to skewed class distribution or imbalanced datasets. Given this potential issue, the exclusive use of testing accuracy rate to measure model performance may lead to biased evaluations. This is because a high test accuracy rate could still be obtained even if the model's proficiency in predicting the minority class is substantially deficient. To provide a more comprehensive evaluation of model performance, we incorporated additional metrics of precision, recall, and the F1 score, alongside test accuracy.

Precision quantifies the ratio of true positive predictions (correct predictions of the "high-level" class in this study) to all positive predictions. A high precision value suggests that the model seldom misclassifies a negative instance as positive. Conversely, **recall** calculates the ratio of true positive predictions to all actual positive instances. A model with high recall excels at detecting positive instances, which is our study's "high-level" students. The **F1** score, the harmonic mean of precision and recall, balances these two metrics. It becomes particularly advantageous when a balance between precision and recall is sought and the dataset exhibits an uneven class distribution.

Although testing accuracy is included as an overall metric to evaluate the model's performance, it is not reported for individual classes because accuracy inherently reflects performance across all classes combined. Reporting class-specific accuracy would be redundant, as it is equivalent to recall for each class and does not provide additional insights. Instead, class-specific metrics such as precision, recall, and the F1 score are used to comprehensively evaluate the model's effectiveness for each class.

For precision, recall, and F1 score, in addition to overall performance evaluation metrics, we also included class-specific performance metrics for both the "high" and "low" levels. This approach provides a more detailed understanding of how well the model performs for each specific class, allowing for a fairer evaluation and evidence-based insights to refine the model.

Providing a more nuanced understanding of model effectiveness is crucial in educational contexts, where the consequences of misclassification can be significantly different for each class. For example, failing to identify students who need additional support (low-level class) could be more detrimental than incorrectly classifying a high-performing student (Bird et al., 2021). With comprehensive metrics, stakeholders (educators, administrators, and technologists) can have greater confidence in the model's deployment, knowing that its performance has been thoroughly evaluated in nuanced ways that reflect the complex realities of educational outcomes. Summarily, this approach supports more informed decision-making in educational contexts where the stakes of accurate and equitable student assessment are high.

The model's performance, assessed using multiple metrics under various scenarios, is detailed in Table 5.

Table 5 provides information facilitating several insights. Primarily, examining learning outcomes associated with water science knowledge in units 2, 3, 4 and 5 reveals a trend where F1 scores are lower than the corresponding test accuracies (refer to the last column of Table 4). On the contrary, for the learning outcomes under the scope of the whole game, related to both water science knowledge and scientific argumentation skills, F1 scores exceed the corresponding test accuracy rates. Such a difference indicates that within the scenarios of the water science knowledge for units 2, 3, 4, and 5, the model excels in predicting the majority class or high-performing group while presenting a weaker capability of predicting the minority class or low-performing group.

Table 5: All data across the metrics are calculated based on the feature set that went through all processes described in the pipeline. This table presents a comprehensive assessment of the model's effectiveness, including Overall F1-score, precision, and recall, alongside detailed metrics for high- and low-performing student groups. The "Overall Performance" section reports F1-score, precision, and recall for the entire student group, while the "High Performance" and "Low Performance" sections provide the same metrics specifically for the high- and low-performing student groups, respectively. These metrics collectively illustrate the model's capability to predict and distinguish between different levels of student performance accurately.

	Overall Performance			High Performance			Low Performance		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Unit 2	84.1% (0.04)	84.7% (0.036)	84.1% (0.042)	86.4% (0.028)	86.3% (0.064)	86.9% (0.049)	81.5% (0.052)	82.5% (0.051)	81% (0.1)
Unit 3	78% (0.035)	81.3% (0.031)	76.9% (0.047)	87.6% (0.015)	85.1% (0.05)	91.3% (0.057)	68.5% (0.06)	77.5% (0.072)	63% (0.14)
Unit 4	85.4% (0.042)	86.9% (0.044)	85.1% (0.046)	90.9% (0.026)	89.7% (0.04)	91.5% (0.059)	79.3% (0.052)	84.3% (0.089)	77.5% (0.105)
Unit 5	91.6% (0.056)	94.9% (0.034)	90.3% (0.074)	95.8% (0.023)	93.9% (0.054)	98.3% (0.024)	86.9% (0.091)	95.6% (0.067)	82.1% (0.164)
Whole Game	88.7% (0.045)	85.9% (0.032)	85.62% (0.033)	88.3% (0.025)	88.2% (0.025)	88.4% (0.057)	82.5% (0.04)	83.6% (0.05)	83.5% (0.05)
Argumentation	87% (0.031)	84.89% (0.032)	85% (0.03)	82% (0.048)	85.89% (0.047)	86.22% (0.031)	84.75% (0.032)	85% (0.032)	83.33% (0.033)

This finding is corroborated by analyzing the corresponding precision and recall values, which show noticeably higher precision than recall. However, a shift is observed when considering the scope of the whole game, including both water science knowledge and scientific argumentation skill, in which the model shows a nearly balanced or closed performance in predicting both "High" and "Low" classes. This finding is confirmed by corresponding overall precision and recall values where recall equals or exceeds precision.

Further validation of the above findings can be achieved by investigating the class-specific F1 score, precision, and recall metrics. The data clearly demonstrates that for water science knowledge in units 2, 3, 4, and 5, all measurement metrics for the high-level class are markedly superior to those for the low-level class; the largest extent can be seen in unit 3. However, the disparity between these metrics for the high-level class and their counterparts for the low-level class is almost negligible when considering water science knowledge and scientific argumentation skill under the scope of the whole game.

Furthermore, Table 5 reveals various model performances in predicting learning outcomes regarding water science knowledge across different units. The model is most effective at predicting learning outcomes for Unit 5 and least effective for Unit 3. By investigating the class-specific metrics in corresponding units, we think the model's limited capability in predicting the learning outcomes of low-level classes is the primary factor contributing to the comparatively weak predictive performance for unit 3. This finding is reinforced by the fact that all metrics specific to the low class in unit 3 fall below the 80% threshold and are significantly lower than their high-class counterparts. Similarly, a notable performance gap between the low and high classes is observed in the water science knowledge of unit 4. Notably, when concentrating on the performance metrics associated with the low-level class, the recall values consistently lag behind precision values, especially in units 3, 4, and 5. Although this observation implies a high accuracy when the model labels a student as a low-level learner, the model's ability to identify low-level learners remains underwhelming. It overlooks several genuinely struggling students who could benefit from additional learning support.

When we focus on the overall learning outcomes of water science knowledge and scientific argumentation skills, the overall performance metrics, as well as the performance metrics for the high-level class of water science knowledge, exceed that of the scientific argumentation skill. Conversely, the predictive performance for a low-level water science knowledge class is similar to the counterpart of the scientific argumentation skill.

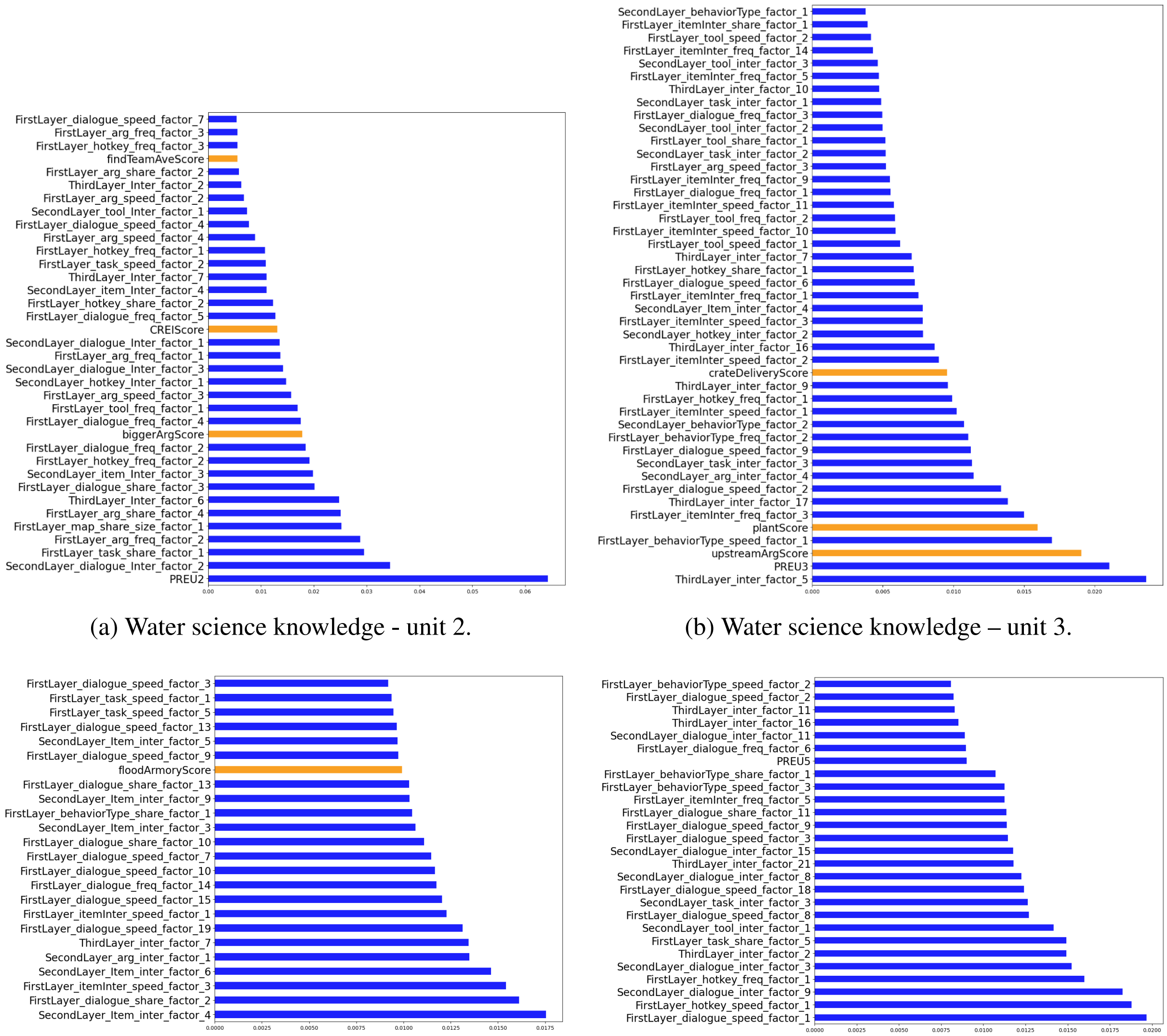
4.2. Model Inference

After a comprehensive evaluation of the model performance metrics, which reflect the effectiveness of our proposed analytic pipeline, we also conducted model inference to examine how involved features operationalized to predict the targeted learning outcomes for understanding their specific impacts. This involved calculating the importance rate for each feature that successfully passed the feature selection process, as we described in section 3.3.12, and creating plots of accumulated local effects (ALE) for expert-crafted features, which represent students' learning progress in different gaming stages and were also retained after the feature selection process. We think these plots provided an in-depth interpretation of the model's outcomes.

4.2.1. Feature Importance Rate

As we mentioned previously, we utilized an ensemble learning model with the hard voting scheme, which involves multiple classification model algorithms, each of which in the ensemble votes for a class ("High" or "Low"), and the class that gets the majority of the votes is chosen as the final prediction. Although this approach is robust against overfitting, one of its issues is that when examining each feature's importance rate, not all algorithms involved within the ensemble model provide a built-in method to calculate features' importance rates. Given the hard voting scheme, which considers all involved classifiers with equal weight, we decided to use the importance of the permutation feature across all classifiers. This method can be applied to any classifier that provides a predictive performance metric and a consistent and model-agnostic metric for assessing feature importance across a diverse array of classifiers within the ensemble model, enabling a holistic understanding of feature contributions devoid of the biases typically associated with model-specific importance measures (Breiman, 2001; Altmann et al., 2010).

Here is a brief illustration of the process we followed to calculate the permutation importance rate for each selected feature. Upon successful training and validation of the ensemble model, the importance of the permutation feature was computed to determine the relative significance of each feature across the model's predictions. This process involved systematically shuffling each feature in the test dataset while maintaining the integrity of all other features, thereby isolating the impact of the shuffled feature on the model's accuracy and stability. The model's performance was evaluated both before and after the permutation of the shuffled feature. For each iteration, the performance metric, which in the current study is the test accuracy, was recorded. The difference in performance metrics quantified the impact of the feature's permutation, indicating its importance. To ensure the reliability and reproducibility of the importance scores, this permutation process was repeated ten times for each feature per dataset split random seed. After the total of 100 (10 repeats * 10 distinct random seeds) repeats, the resultant changes in performance were then averaged to mitigate the effects of random variations inherent in the permutation process.



(a) Water science knowledge - unit 2.

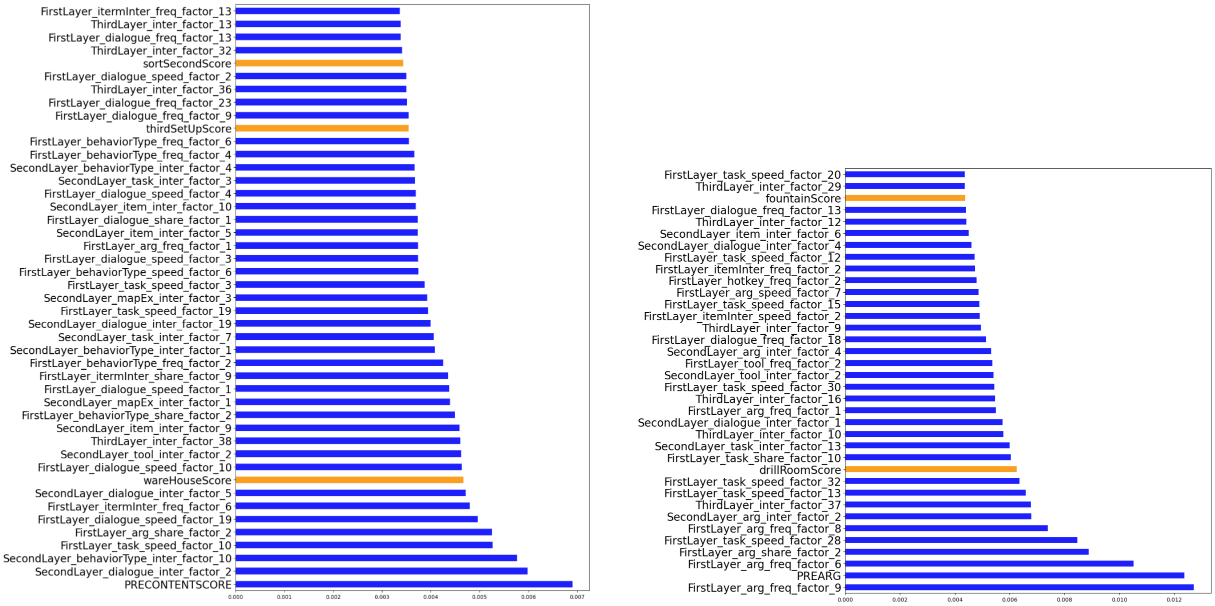
(b) Water science knowledge – unit 3.

(c) Water science knowledge – unit 4.

(d) Water science knowledge – unit 5.

Figure 3: Ordered Permutation Importance Rate Across Game Contexts. Orange bars represent learning progress features. Blue bars represent other features.

After an initial glance at Figures 3 and 4, we can summarize some key findings to provide a clearer context for the subsequent analysis: 1) the dialogue-related behaviors contribute largely to the learning outcome prediction of the water science knowledge in units 2, 3, 4 and 5, across the whole game and the scientific argumentation skill; 2) The argumentation-related behaviors significantly contribute to learning prediction of the overall scientific argumentation skill and the water science knowledge of unit 2; 3) For the item-interaction behaviors, we can see it



(a) Water science knowledge – whole game.

(b) Scientific argumentation skill – whole game.

Figure 4: Ordered Permutation Importance Rate Across Game Contexts. Following Figure 3.

contributes significantly to the learning prediction in regards to the water science knowledge in units 3, 4, and the overall game scale; 4) Regarding the tool usage behaviors, they contribute largely to the learning outcome prediction in units 3; 5) The task completion behaviors have significant contribution to the learning prediction regarding overall water science knowledge and scientific argumentation skill; 6) The shares of different behavior types (events) contributes to predicting the learning outcomes of the water science knowledge of unit 5 and the whole game scale. These findings set the stage for a deeper analysis of the underlying contributions of these behaviors in various game contexts.

Building on these general findings, we examine the specific contributions of in-game features to learning outcome predictions in each game context. Figures 3 and 4 provide detailed results of the permutation importance rates for each context. Our first inquiry focuses on what types of in-game actions significantly contribute to the learning outcome prediction under each context. We first discovered that contributions of 8%, 13%, 4%, 15%, 9%, and 17% are attributed to intersectional features across all behavior types (the features' names start with "ThirdLayer") in units 2, 3, 4, 5, the water science knowledge for the whole game, and the scientific argumentation skill for the entire game, respectively. Interestingly, we infer that at least 80% of the contributions are attributed to features representing a specific in-game behavior type (combining features with names starting with "FirstLayer" and "SecondLayer" under a certain behavior type) across different game contexts.

A closer look at individual units reveals further details about these contributions. Within Unit 2 (Figure 3a), dialogue- and argumentation-related behaviors account for 25% and 22% of the predictive contribution among individual behavior types. In unit 3 (Figure 3b), 29% of predictive performance originates from item interaction behaviors, 13% is derived from dialogue reading behaviors, and 11% stems from tool usage behaviors among individual behavior types. Concerning unit 4 (Figure 3c), dialogue-related attributes contribute 46%, and item interaction behaviors contribute 29% within individual behavior types. In unit 5 (Figure 3d), the primary contribution comes from dialogue-related behaviors, with 48% shares among individual behavior types, followed by behavior type shares, with shares of 11%. Under the whole game scope (Figure 4a), for water science knowledge, the dialogue-related behaviors lead, contributing 27% among individual behavior types, followed by item interaction behaviors contributing 18%, shares of different behaviors (events) contributing 16%, and task-completion behaviors with 13%. Regarding argumentation skills (Figure 4b), argumentation-related behaviors contribute the most, with 28% shares. The second most contributed behavior type is dialogue-related behaviors, with 25% shares, followed by task-completion behaviors, with 22% shares.

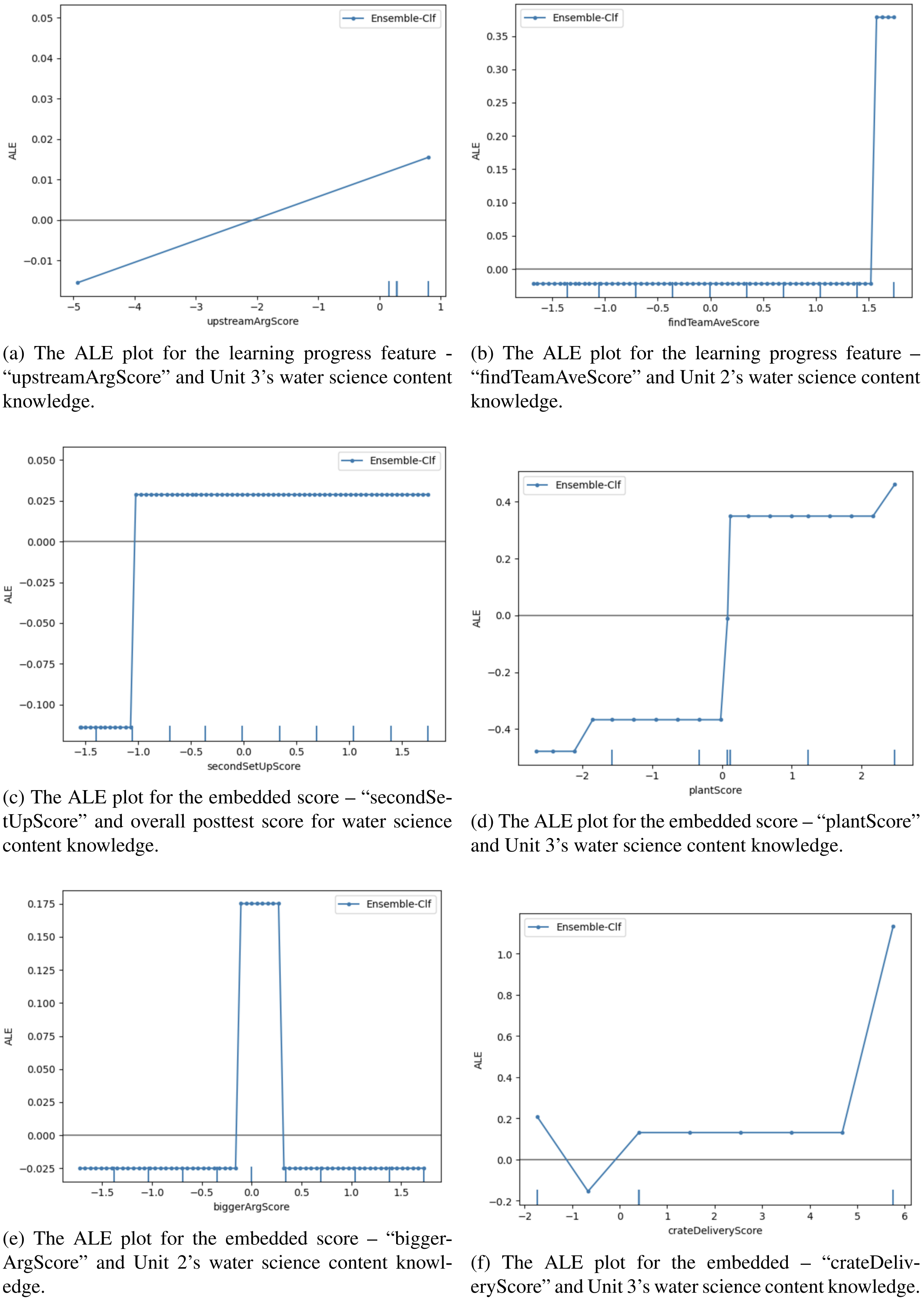
In addition to these behavior-specific contributions, certain predictors stand out as highly influential. For instance, in some game contexts, pre-assessment scores emerge as the most important predictor, such as for water science knowledge in unit 2 and the overall game scale. However, for water science knowledge in unit 4 (Figure 3c), the corresponding pre-assessment score is excluded from the selected features. Additionally, bar plots in Figures 3 and 4 highlight the importance of features representing students' learning progress at different game stages. Specifically, these include three features for Unit 2, three for Unit 3, one for Unit 4, three for water science knowledge across the whole game, and two for scientific argumentation skills.

There are 12 learning progress features involved within prediction models across different game contexts, and we are interested in studying the impact of each expert-crafted feature on the final learning outcome per context. This pursuit is motivated by gaining comprehensive insights into questions such as how our learning progress features, which quantitatively assess students' progress or milestones, influence the prediction of their post-assessment learning outcomes, and what perspectives we can extract from interpreting the nature of the influence to refine the game and pedagogical design.

4.2.2. Feature Interpretation: Accumulated Local Effects Plots

In our previous description, we mentioned that we used an ensemble learning algorithm with the hard-voting scheme, which includes ten base classifiers to predict the targeted learning outcome, given our complex and adaptive feature sets. Most of these base classifiers are characterized as "black-box" supervised learning models. A significant drawback of these models is their lack of interpretability or transparency, making it challenging to examine how the included predictors influence the predicted responses. Applying partial dependence (PD) plots to estimate feature effects is a common strategy to overcome this shortcoming. However, PD plots require the tested features to be uncorrelated, which is hard to meet in datasets collected within real-world contexts. As a viable alternative, accumulated local effects (ALE) can handle features that exhibit correlation with one another and provide visualizations depicting how the changes in each feature's value impact the probability of class classification of the dependent variable (Apley and Zhu, 2020).

For several reasons, we draw ALE plots exclusively for features representing students' learning progress, as selected by the final ensemble learning model. Firstly, these features are expert-crafted, inherently interpretable, and meaningful within the educational domain. This facilitates a clearer understanding of how variations in these features influence the model's predictions. Secondly, the interpretability of these features is crucial for deriving actionable insights and effectively communicating results to stakeholders, particularly non-technical ones. Thirdly, en-



(a) The ALE plot for the learning progress feature - “upstreamArgScore” and Unit 3’s water science content knowledge.

(b) The ALE plot for the learning progress feature – “findTeamAveScore” and Unit 2’s water science content knowledge.

(c) The ALE plot for the embedded score – “secondSetUpScore” and overall posttest score for water science content knowledge.

(d) The ALE plot for the embedded score – “plantScore” and Unit 3’s water science content knowledge.

(e) The ALE plot for the embedded score – “biggerArgScore” and Unit 2’s water science content knowledge.

(f) The ALE plot for the embedded – “crateDeliveryScore” and Unit 3’s water science content knowledge.

Figure 5: ALE plots for standardized scores measuring students’ in-game learning progress

semble learning models' selection of these features indicates their statistically significant influence on model predictions. In contrast, other features, which are latent and generated by feature extraction techniques representing students' in-game behaviors, often lack straightforward interpretations, diminishing their utility in explanatory analyses. Therefore, focusing on features that represent learning progress enhances the relevance and comprehensibility of the model's results, ensuring that the analysis remains accessible and meaningful to a broader audience.

Figure 5 presents parts of the ALE plots, which display representative or typical shapes or trends between learning progress features and their corresponding learning outcomes. Appendix F contains ALE plots and detailed illustrations for all selected learning progress features. Understanding the ALE plots can be aided by the subsequent explications of the key involved components:

1. **X-axis:** Represents the range of standardized feature values used in the study.

2. **Y-axis:** Indicates the variation in the predicted probability of the positive class (in our study, the high-level class) as the feature value changes.

3. **Horizontal line at y = 0:** serves as a reference point. An ALE curve crossing this line suggests that feature values below it decrease the predicted probability of the positive class, whereas values above it increase the probability (or vice versa).

4. **Shape of the plot:** The ALE plot's shape reveals the relationship between the feature and the predicted probability, which can be linear, non-linear, non-monotonic, or exhibit complex patterns.

5. **Magnitude of ALE value:** Reflects the importance of the feature; larger magnitudes indicate a more significant impact on the predicted probability.

Their relative orientation remains consistent with the original scores despite standardizing score values. This means the transformed values' lower range (leftmost side) corresponds to the lower range of the original scores, and the upper range (rightmost side) corresponds to the upper range of the original scores. To facilitate interpretation, we correlate student in-game performance metrics with the respective intervals of the standardized score values, as detailed in Table 6.

Table 6: Score intervals and corresponding in-game performances for selected features representing students' learning progress.

Feature Name and Brief Illustration	**Standardized score value interval**	**In-game performance**
(a) UpstreamArgScore: Measures progress in constructing a scientific argument to identify the pollutant source in Unit 3.	Leftmost interval	Failed the task or made more than six attempts without a correct argument.
	Middle interval	Submitted a correct argument within three attempts.
	Rightmost interval	Correct argument submitted on the first attempt.

Table 6: Score intervals and corresponding in-game performances for selected features representing students' learning progress.

Feature Name and Brief Illustration	**Standardized score value interval**	**In-game performance**
(b) findTeamAveScore: Measures progress in locating the team using topographic knowledge in Unit 2.	Leftmost interval	Abandoned the task or found the team after three minutes without using the topographic map.
	Middle interval	Found the team after more than three minutes using the map, or within three minutes without using the map.
	Rightmost interval	Successfully found the team within three minutes, demonstrating strong topographic map skills.
(c) secondSetUpScore: Measures progress in selecting the correct machine to generate distilled water on the second floor of a factory in Unit 5.	Leftmost interval	Failed to complete the task or selected the wrong machine type.
	Middle interval	Eventually selected the correct machine after several attempts.
	Rightmost interval	Chose the correct machine type on the first attempt.
(d) plantScore: Measures progress in selecting correct seedbeds to install pumps along a river, assessing knowledge of dissolvable materials in Unit 3.	Leftmost interval	Failed to complete the task or installed all pumps in incorrect seedbeds.
	Middle interval	Installed some pumps correctly, but others in incorrect seedbeds.
	Rightmost interval	Successfully installed all pumps in the correct seedbeds, demonstrating complete accuracy.
(e) biggerArgScore: Measures progress in constructing a scientific argument to determine the larger watershed in Unit 2.	Leftmost interval	Disengaged from the task without submitting a correct argument.
	Middle interval	Submitted a correct argument within three attempts.
	Rightmost interval	Submitted a correct argument on the first attempt.
(f) crateDeliveryScore: Measures progress in delivering crates to an NPC's base based on water flow direction.	Leftmost interval	Failed to complete the task or delivered all crates via the wrong river.
	Middle interval	Delivered some crates correctly, but others via the wrong river.
	Rightmost interval	Successfully delivered all crates via the correct river, demonstrating complete accuracy.

ANALYSIS OF ALE PLOTS DISPLAYING INCREASING TRENDS The ALE plots in this subsection exhibit approximately increasing trends between students' standardized scores representing their learning progress and corresponding learning outcomes. However, variations exist among these plots, with one showing a linear relationship, another showing a steady non-linear increase, and others demonstrating non-linear relationships with abrupt increases.

For the "upstreamArgScore" – which assesses students' performance in the scientific argumentation task in Unit 3 – a clear linear relationship is evident between the standardized score and the likelihood of being classified as a high-level learner, as illustrated in Figure 5a. This task requires students to persuade a character, Bill, about the source of pollution based on evidence. As shown in row (a) of Table 6, students who present a correct argument with fewer attempts are likelier to achieve high-level learner status in the posttest.

In the case of the "plantScore" – which measures students' progress in installing pumps for garden beds downstream of a large polluted tree based on their understanding of dissolvable materials in water – the ALE plot in Figure 5d reveals a non-linear but steadily increasing impact on performance. As noted in row (d) of Table 6, students incorrectly placing more than two pumps are less likely to be classified as high-level learners. In comparison, correctly installing more than two pumps significantly boosts this likelihood.

Figures 5b and 5c display ALE plots with non-linear relationships characterized by sudden increases at different score points. Figure 5b shows this sudden change at the rightmost score interval, while Figure 5c shows it at the leftmost score interval. Specifically, Figure 5b examines the relationship between post-test scores and students' performance in locating their teams. Row (b) of Table 6 indicates a significant positive impact on posttest scores for the highest performance range, with ALE values exceeding the zero threshold. Students who locate their team within three minutes and effectively use the topographic map are significantly more likely to attain a high-level understanding of Unit 2's water science content. In contrast, other performance levels are associated with lower post-test outcomes, suggesting a greater likelihood of achieving only a low-level understanding of the material. Learning progress features which present similar trends to Figure 5b ("findTeamAveScore"), include "floodArmoryScore," "wareHouseScore," and "fountainScore."

Regarding Figure 5c, which examines the relationship between the "secondSetUpScore" and the overall posttest score for water science knowledge, row (c) of Table 6 indicates a positive correlation between students' posttest scores and their performance in this task, provided they select the correct machine, regardless of the number of attempts. Conversely, incorrect machine selection is associated with lower post-test scores. The learning progress feature "thirdSetUpScore" has a similar trend to "secondSetUpScore."

Analysis of ALE Plots Displaying Bell-Like Shape

The bell-like shape in ALE plots is characterized by negative impacts on learning outcomes at the leftmost and rightmost score intervals, while the middle interval has a positive effect. Figure 5e exemplifies this shape. It illustrates the relationship between posttest scores measuring students' water science knowledge and their performance in scientific argumentation within Unit 2. The plot reveals that the lowest and highest score intervals negatively affect the likelihood of attaining a high-level classification in the corresponding learning outcome. In contrast, an intermediate score range of -0.2 to 0.3 positively influences the probability of achieving a high-level classification. As shown in row (e) of Table 6, students who submit correct arguments within three attempts but not on the first attempt are likelier to achieve high-level outcomes as measured by the posttest.

Analysis of ALE Plots Displaying Complex and Oscillating Patterns The ALE plot selected for this subsection exhibits a complex and oscillating trend, which is more challenging to interpret than the other plots.

Figure 5f shows the ALE plot for "crateDeliveryScore," which evaluates the accuracy of delivering crates based on water flow in a specific task. Combined with row (f) of Table 6, the plot indicates that delivering three crates incorrectly while delivering one correctly negatively impacts post-test scores. Conversely, correct delivery—mainly when all crates are delivered accurately—correlates positively with higher post-test scores. This pattern underscores the importance of accuracy in this task for better knowledge acquisition in Unit 3's water science content. The challenging aspect of interpreting this ALE plot lies in explaining the observed phenomenon where no correct delivery still positively impacts post-test scores. This may require higher-dimensional ALE plots to provide a more straightforward explanation. Learning progress features that show similar trends to "crateDeliveryScore" include "CREiScore" and "drillRoomScore."

Summary Among the selected learning progress features, only "UpstreamArgScore" demonstrates a linear relationship with the probability of classification into high-level learning outcomes, as measured by posttest assessment scores. The other features exhibit non-linear relationships. Not all features show a consistent positive correlation, indicating that better in-game performance does not always correspond to better post-game learning outcomes. This is evident in learning progress features such as "biggerArgScore," and "crateDeliveryScore." The ALE plots for these features reveal complex, non-linear, and oscillating patterns, suggesting a need for further analysis to clarify these relationships.

4.3. Contribution Summary

Before beginning our discussion, it is essential to revisit the foundational research questions that have guided our study. These questions have provided direction and focus, supporting us in describing and validating the proposed analytic pipeline and interpreting its results.

4.3.1. RQ1: What Distinct Elements are Encapsulated within the Overarching Analytic Pipeline?

In addressing RQ1, which aimed to identify the distinct elements within our analytic pipeline, we outlined a methodical 16-step process. The pipeline begins with extracting raw logs from the game logging system, followed by comprehensive cleaning and categorization to ensure data integrity. Key stages involve the extraction and standardization of both learning progress features and in-game behavior features. The in-game behavior features were further categorized to provide a detailed understanding of player interactions.

We employed unsupervised learning techniques to improve feature clarity, reduce noise, and optimize predictive accuracy while minimizing information loss. These techniques enabled dimensionality reduction and the identification of latent variables within individuals and across multiple behavior types. The result was a set of refined features that captured intricate learning patterns and the interplay between different behaviors, offering a comprehensive view of student interactions and learning outcomes.

During the modeling phase, consistent data preprocessing and standardization were prioritized. We adopted an ensemble learning approach, utilizing 10 diverse, robust, and complementary classifiers. This strategy was designed to address our feature set's complexities and dynamic nature, ensuring robust and generalized predictions across various game contexts and outcomes for different learning objectives. Additionally, we conducted a feature selection process during model training to reduce noise and improve the signal within the feature set.

To validate the effectiveness of the analytic pipeline, we employed multiple model performance metrics using feature sets generated at different stages of the pipeline. These metrics were evaluated against rich gameplay logs collected from the DGBL environment Mission HydroSci (MHS). The results confirmed the pipeline's ability to significantly enhance prediction model performance across different game contexts and learning subjects, demonstrating its potential applicability to other DGBL environments. A more detailed summary of the pipeline's validation is provided in the subsequent subsection.

4.3.2. RQ2: How Effective is This Pipeline across Various Contexts?

To evaluate the efficacy of the analytic pipeline across various contexts within the MHS environment, we systematically assessed its ability to predict different learning outcomes. The analysis of testing accuracy rates derived from feature sets at various pipeline stages, confirmed that the comprehensive feature processing significantly enhances model predictive accuracy. The final, post-pipeline feature set consistently improved testing accuracy across all scenarios, surpassing the 80% threshold acceptable in educational settings. This outcome highlights the critical role of advanced feature engineering techniques in capturing the complex patterns of student interactions and learning outcomes, such as dimensionality reduction and the integration of intersectional features.

Our analysis also revealed variability in model performance across different game contexts (units) and learning objectives. When comparing different knowledge domains within the entire game context, the accuracy for overall content knowledge in water science slightly exceeded that for scientific argumentation. Notably, the pipeline was most effective in predicting learning outcomes for water science knowledge in Unit 5 and least effective in Unit 3.

The lower accuracy observed in Unit 3 could be attributed to the increased complexity and novelty of its content and mechanics, which were new and more diverse for students than earlier units. Unlike Unit 2, where content variations reappear in later units, Unit 3's unique challenges likely demanded more cognitive resources from students, making it harder for them to solve in-game puzzles and tasks using prior knowledge. This may have led to frustration, confusion, and disengagement, ultimately impacting their performance.

In contrast, the higher accuracy observed in Unit 5 may be due to the consistency of core game mechanics—such as dungeon puzzles, item interactions, and environmental exploration—despite differences in specific mechanics. Familiarity with these core mechanics likely helped students maintain their progression in the game and focus on mastering the targeted content knowledge and skills. This increased their likelihood of being classified into the high-level learning outcome category and mitigated the model's difficulty in predicting the minority class.

The inclusion of precision, recall, and F1 score metrics provided a more nuanced evaluation of model performance, particularly in predicting high and low learning classes separately. While the models excelled in predicting high-performing students, their effectiveness in identifying low-performing students was less robust, particularly within individual units. This shortcoming likely stems from the absence of features specifically capturing off-task behavior, disengagement, and guessing. The inability to measure these factors reduced the accuracy of predicting low-level learning outcomes. This finding highlights the need for future research to improve model performance in predicting low-performing students, as misclassification in this group can have more severe consequences in educational contexts.

4.3.3. RQ3: What Methods Can Help Interpret the Black-box Computational Models, and What Insights Can be Drawn from their Results?

To address RQ3, we explored methods for interpreting black-box computational models, focusing on enhancing model transparency and understanding. We employed two primary techniques: permutation feature importance and ALE plots.

PERMUTATION FEATURE IMPORTANCE Through permutation importance analysis, we identified that both individual in-game behaviors—such as dialogue-related, argumentation-related, item interaction, tool usage, and task completion behaviors—and intersectional features that combine different behavior types consistently play a crucial role in predicting learning outcomes across various game contexts. This finding highlights the need to consider specific behaviors and their interactions when designing predictive models in educational settings.

The analysis also revealed that the types of behaviors most influential in predicting learning outcomes vary across different game contexts. This finding suggests that the design of each DGBL environment strongly influences which behaviors become significant predictors of the targeted learning objectives. For example, in Unit 2, extensive dialogues between the main character and non-player characters (NPCs) convey critical background information, enhance engagement, and guide quest completion. As a result, dialogue-related behaviors emerge as the most influential in model construction. In contrast, Unit 3 emphasizes puzzles requiring interaction with various in-game items, making item interaction behaviors the most critical for predicting learning outcomes related to Unit 3's content knowledge.

Additionally, our analysis revealed variability in the importance of pre-assessment scores across different scenarios. In certain units, pre-assessment scores were not the most critical predictors, suggesting that other in-game behaviors may have a more direct impact on learning outcomes. This underscores the importance of context-specific feature selection depending on the game environment and learning objectives. Especially for unit 4's water science knowledge, the pretest score is absent in the final prediction model, although it correlates significantly with its posttest score. This could suggest that the pedagogical and game design of Unit 4 effectively facilitate students' mastery of the targeted learning objectives within the game, reducing their reliance on prior knowledge to achieve specific outcomes.

ACCUMULATED LOCAL EFFECTS (ALE) PLOTS We further emphasized the value of expert-crafted features representing students' learning progress at different game stages. The model selected these features as important predictors and we analyzed their influence using ALE plots. This approach revealed various relationships between these features and learning outcomes. For instance, while some features like "UpstreamArgScore" displayed a clear linear relationship with high-level learning outcomes, others, such as "crateDeliveryScore," exhibited more complex, non-linear patterns. These findings indicate that in-game performance does not always straightforwardly translate to better learning outcomes, highlighting the complexity of these relationships and the need for sophisticated analytical methods to fully understand them.

The use of ALE plots contributes to the broader goal of enhancing model transparency in educational data mining. By providing clear visualizations and interpretations of feature impacts, the research makes predictive models more accessible and usable for stakeholders, including educators and administrators. This transparency is crucial for ensuring that predictive models are not only accurate but also comprehensible and actionable in real-world educational settings. Additionally, the insights derived from ALE plots have practical implications for designing educational games and developing pedagogical strategies. Educators and game designers can refine their approaches to better support student learning by understanding how specific in-game behaviors and learning progress features influence outcomes. The findings offer a foundation for improving both game mechanics and educational interventions, with specific recommendations and potential refinements discussed in the following section.

5. Discussion

5.1. Further Comprehension of ALE Plots

The analysis of ALE plots identified 12 learning progress features that significantly impacted the targeted learning objectives. However, these features exhibited different patterns in influencing outcomes as their values changed. Broadly, these patterns can be categorized into three groups: "Increasing Trend," "Bell-like Shape," and "Complex and Oscillating Patterns," as discussed in Section 4.2.2.

5.1.1. Increasing Trend

This group includes most selected learning progress features—8 out of 12 fall into this category. Upon closer inspection, these features can be further divided into three subgroups: "Threshold Incremental Learning," "Progressive Learning Gain," and "Learning Inflection Point."

Threshold Incremental Learning ALE plots in this subgroup display a distinctive pattern: the leftmost range falls below zero on the y-axis, while the middle and rightmost ranges rise above it. This suggests an overall increasing trend in student performance on these learning progress values, particularly after surpassing a specific threshold. Each quest associated with these features incorporates feedback elements, leading us to hypothesize that the feedback mechanisms within MHS—such as pop-up dialogues, auditory feedback from an AI robot, and changes in game scenes based on student choices—significantly influence these patterns. Students who fail to adapt their strategies in response to feedback tend to display lower learning outcomes, while those who effectively use feedback to adjust their decision-making are more likely to be classified as high-level learners.

Constructivist theory provides an alternative explanation, positing that learners construct knowledge through interactions with their environment and reflection on those experiences (Pivec et al., 2003; Bada and Olusegun, 2015). In the MHS environment, students learn from actions such as submitting arguments, assembling machine parts, and using topographic maps, adjusting their understanding based on feedback. As indicated by the ALE plots, students who do not adapt their strategies (leftmost range) are less likely to achieve high learning outcomes, while those who actively learn from feedback (middle and rightmost ranges) are more likely to succeed.

Progressive Learning Gain This subgroup includes features that show an ascending trend, varying only in linearity. The ALE plots for these features indicate a positive correlation between the learning progress scores and the ALE values—more accurate decisions and fewer errors correspond to improved learning outcomes.

While constructivist theory explains some aspects, operant conditioning theory (Staddon and Cerutti, 2003; Akpan, 2020) may better account for scenarios where the impact on learning outcomes decreases with more correct choices or where the middle range of learning progress values has a negligible effect. Positive reinforcement (Staddon and Cerutti, 2003; Kirsch et al., 2004; Wu et al., 2012) likely plays a role, as students who receive positive feedback after correct decisions are more motivated to continue using successful strategies, thereby enhancing learning outcomes. Conversely, punishment or negative reinforcement following incorrect choices prompts students to adjust their strategies, improving their performance.

Cognitive load theory (Sweller, 1994; Sweller, 2011) also offers insights, suggesting that the leftmost range may represent a high cognitive load where students struggle to assimilate new information, leading to underperformance. The rightmost range may indicate effective cognitive load management, where students have acquired the necessary knowledge and skills, improving their problem-solving abilities. The middle range could represent a transitional stage where students gradually improve their cognitive load efficiency.

Learning Inflection Point In this subgroup, ALE plots reveal that the leftmost and middle intervals fall below zero, while the rightmost range rises above it. This suggests that only students with high learning progress values will likely be classified as high-level learners. Mastery learning theory (Block and Burns, 1976; Slavin, 1987; Yang, 2017; McGaghie and Harris, 2018) provides a cogent framework for interpreting these patterns. This theory posits that students must master one topic before progressing to the next. Students who make the correct choice on their first attempt demonstrate mastery of the necessary skills and knowledge. At the same time, those who struggle may need additional support or interventions to reach mastery. Cognitive load theory also offers an explanation. High-level learners likely manage their cognitive load effectively, balancing the complexity of tasks with their available cognitive resources. Students who struggle may be experiencing increased cognitive load due to complex instructions or problem-solving demands, and this higher load can impede learning.

5.1.2. Bell-Like Shape

In this category, ALE plots resemble an arch, with both extremes falling below zero and the midpoint rising above it. This pattern suggests that extremely low or high values on these learning progress features correlate with lower learning outcomes, while intermediate scores yield better outcomes. This effect may be due to MHS's integrated formative feedback approach. Students who fail to make correct decisions may disengage, leading to frustration and missed learning opportunities. Conversely, students who make correct decisions too quickly may rely too heavily on pre-existing knowledge, missing new insights. Those who adapt their choices based on feedback tend to enhance their understanding and performance, aligning with Vygotsky's Zone of Proximal Development theory, which suggests optimal learning occurs when tasks are neither easy nor difficult (Chaiklin et al., 2003).

5.1.3. Complex and Oscillating Patterns

Features in this group generate ALE plots with unstructured and oscillatory patterns, with some intervals above and others below zero. While certain intervals align with expectations—such as higher scores increasing the likelihood of high learning outcomes—other unexpected patterns emerge, making these ALE shapes difficult to explain.

Cognitive load theory offers a potential explanation for some of these patterns. Students in the rightmost ranges may manage their cognitive load effectively, leading to better outcomes, while those in the leftmost ranges may struggle with higher cognitive loads, impairing their learning. The oscillatory nature of the middle interval suggests a complex interplay of cognitive loads, where learning outcomes do not change linearly but fluctuate based on task complexity, prior knowledge, and instructional effectiveness.

These findings indicate that further research, possibly involving higher-order ALE plots, is needed to fully understand the dynamic interactions among these factors.

5.2. Thoughts on Game Design Issues

As discussed in the previous sections, some learning progress features produced ALE shapes that deviated from our initial expectations, particularly those in the “Bell-like Shape” and “Complex and Oscillating Patterns” categories. These unexpected trends suggest potential issues related to the design and implementation of these learning progress features. Three key unexpected trends emerged:

1. Students who achieved high learning progress values (typically by submitting correct answers on their first attempt) were not consistently classified as high-level learners in the corresponding learning outcomes.

2. Students with mid-range learning progress values did not exhibit an increasing likelihood of achieving high-level learner status, compared to those with low-range learning progress values, which is contrary to expectations.

3. Fluctuating and oscillating patterns appeared around the x-axis zero threshold in the ALE plots, which are difficult to interpret.

Several factors may contribute to these trends, including possible shortcomings in how the learning progress features were crafted, which could fail to capture all necessary information to predict the targeted learning outcomes. However, this section will focus on discussing the game design-related issues that might underlie these trends, briefly mentioning potential issues in feature generation.

5.2.1. High Learning Progress Values Not Leading to High-Level Learner Status

One possible explanation for this trend is that the game design of MHS might encourage surface learning, where students focus on getting the correct answer without fully understanding the underlying concepts. Students may rely on memorization or pattern recognition rather than engaging deeply with the content, allowing them to submit correct answers initially but not leading to a thorough comprehension of the material—critical for long-term retention and knowledge transfer.

Additionally, the game may not provide sufficient opportunities for students to integrate new knowledge with existing schemas or practice newly learned concepts in varied scenarios. Without these opportunities, students are less likely to develop the deeper understanding required to achieve high-level learning outcomes.

Another issue might be the limited feedback provided within the game. If MHS fails to encourage students to reflect on their choices or to understand why an answer is correct, students might not fully engage with the learning process. Even when they submit correct answers, they may not grasp the underlying principles or how they apply to broader concepts, leading to a gap between performance on individual tasks and overall learning outcomes.

Finally, there might be a misalignment between cognitive load and task complexity. High progress scores could reflect low cognitive load for certain tasks, meaning students find them easy and do not need to engage deeply to succeed. However, this lack of cognitive engagement might prevent the development of higher-order thinking skills, which are necessary for high-level learning outcomes. Similarly, the tasks associated with high progress values might be too simple or disconnected from more complex, integrative challenges. This simplicity allows students to succeed at a certain stage but does not prepare them for more complex problem-solving, requiring deeper understanding and higher-level skills.

5.2.2. Mid-Range Learning Progress Values Not Correlating with Higher Learner Status

Several game design issues might explain this unexpected trend. First, the feedback provided after incorrect attempts might be insufficient or unclear, preventing students from understanding their mistakes and making necessary adjustments. Without constructive feedback guiding them toward the correct approach, students' learning processes could be hindered, leading to lower learning outcomes despite eventual success.

Second, repeated attempts or incorrect decisions may lead to cognitive overload and fatigue. The effort required to process information repeatedly and attempt different strategies could exhaust students' cognitive resources, resulting in mental fatigue. This fatigue might prevent them from maintaining the concentration and effort needed to achieve high-level learning outcomes.

Third, some in-game tasks might include ambiguous instructions. Students may make mistakes not because they lack understanding but because they misunderstand the task requirements. This could lead to unnecessary errors, reducing motivation and engagement and negatively impacting learning outcomes.

Overall, MHS may benefit from refining its systems or game mechanics to better support students in correcting their mistakes and receiving appropriate rewards for such behaviors.

5.2.3. Fluctuating and Oscillating Patterns in ALE Plots

Several factors could contribute to these fluctuating and oscillating patterns. First, tasks associated with these learning progress features might simultaneously teach or provide information on multiple skills or knowledge areas. If these objectives are not well-integrated or prioritized, progress in one skill might not translate effectively to outcomes in another, creating complex interactions that are difficult to interpret.

Second, MHS might allow for multiple strategies without providing sufficient guidance. Players may adopt various approaches—including luck-based guessing—to achieve objectives or complete tasks. Some strategies might be more effective than others in promoting learning, leading to inconsistent patterns in the data.

Third, better integration of educational content with game mechanics may be needed. If there is a disconnect between learning objectives and game elements, progress in the game may not effectively translate to learning outcomes.

Lastly, the expert-crafted learning progress features might be interdependent or have hidden dependencies with each other or with features representing in-game behaviors. These interdependencies can create complex interactions that are not immediately apparent. Addressing this issue may require additional iterations of feature engineering to refine the dataset used for training the model.

5.3. Advancing DGBL Environments Design Based on Empirical Findings

Building on previously summarized empirical observations and analyses, we have synthesized the following recommendations to guide future advancements in DGBL environment design.

5.3.1. Enhance Scaffolding to Balance Novelty and Familiarity

To address the challenges posed by increasing complexity and novelty within in-game units, practitioners should consider enhancing scaffolding techniques to balance these elements effectively. For units like Unit 3, where both complexity and novelty are heightened, introducing new game mechanics incrementally can facilitate better learning outcomes (Grey et al., 2017). This approach involves beginning with simpler tasks that build foundational understanding before progressing to more complex challenges.

One effective strategy is to utilize familiar game mechanics from previous units to introduce new pedagogical concepts. For example, at the beginning of a complex unit, employing game mechanics that players have already mastered allows them to focus on assimilating new content without the added cognitive load of learning unfamiliar mechanics. Once students have internalized the new concepts through familiar gameplay, new game mechanics can be introduced to enable them to apply these concepts in novel contexts. By combining new content with familiar game mechanics, this method reduces the potential for cognitive overload and leverages prior knowledge, making it easier for students to adapt to new challenges (Braad et al., 2016).

Alternatively, establishing core game mechanics that persist across units while introducing variations—such as the argumentation engine within MHS—can aid in the transfer of skills learned in earlier units to new contexts. This consistency provides a stable framework within which students can navigate new content, promoting deeper understanding and skill retention.

Another valuable approach is to implement tutorials or guided practice sessions before introducing new game mechanics. These sessions focus solely on familiarizing students with the new mechanics without integrating pedagogical or curriculum content initially. Once players become comfortable with the new mechanics, curriculum content can then be seamlessly blended into the gameplay. This staged introduction allows students to concentrate on mastering one aspect at a time, thereby managing cognitive load more effectively.

By adopting these strategies, educators can break down complex information into smaller, manageable chunks, preventing students from becoming overwhelmed. It is crucial to ensure that students focus on one learning objective at a time—either acclimating to new game mechanics or learning new curriculum content—to reduce unnecessary cognitive burden. This refined approach to game design not only enhances the learning experience but also leads to improved engagement and better learning outcomes.

5.3.2. Optimize Feedback Mechanisms

Optimizing feedback mechanisms is crucial for enhancing learning outcomes in DGBL environments. Feedback content should be adapted to students' prior performance to maximize its effectiveness (Johnson et al., 2017). Initially, when students make mistakes in their first or early attempts, providing comparatively generic feedback can encourage them to think more deeply about the problems, fostering better understanding and more accurate responses. As the number of unsuccessful attempts increases, feedback should become more specific and actionable, offering tailored suggestions based on their actions to help them adjust their strategies effectively.

Practitioners should utilize various forms of feedback—visual, auditory, and kinesthetic—to reinforce understanding, maintain student engagement, and cater to different learning styles. The timing of feedback is also a critical factor that should align with the design objectives. If the goal is to reinforce immediate learning, providing instant feedback for in-game actions is beneficial. Conversely, if the aim is to encourage deeper cognitive processing, implementing delayed or reflective feedback can prompt students to engage more thoughtfully with the content.

By carefully designing adaptive, multimodal, and strategically timed feedback mechanisms, educators can enhance the educational effectiveness of DGBL environments. This approach not only supports students in correcting errors but also promotes deeper engagement and long-term retention of the material.

5.3.3. Monitor Engagement Levels to Address Off-Task Behavior and Guessing

To mitigate off-task behavior and ensure sustained engagement and learning, monitoring student actions through the integrated logging system can offer insights for targeted interventions and game adjustments (Biedermann et al., 2023). By analyzing engagement metrics, patterns indicative of disengagement or random guessing—such as rapid sequences of incorrect responses—can be detected. Based on these insights, game mechanics can be designed to mitigate the effectiveness of guessing. For instance, implementing penalties for consecutive incorrect attempts or requiring students to provide justifications for their answers can discourage superficial engagement.

If the system detects that a student is exhibiting guessing behavior, the game could temporarily halt the progression of the main storyline. Instead, the student would be redirected to supplementary side tasks designed to reinforce familiarity with the curriculum content. Upon successful completion of these tasks, the student would be allowed to resume the main storyline. Additionally, implementing a reward system that recognizes consistent engagement and thoughtful responses can further encourage sustained focus.

It is also important to acknowledge that off-task behavior or guessing may be symptomatic of fatigue. In such cases, enhancing in-game task design to ensure that tasks are challenging yet not overwhelming is crucial (Kanal et al., 2020). Incorporating moments of rest or lighter activities within the game can help prevent cognitive overload and maintain student engagement. By balancing task difficulty with opportunities for recuperation, practitioners can create a more sustainable and effective learning environment within the game.

5.3.4. Foster Consistent Motivation

Maintaining student motivation throughout gameplay is key to effective learning. The following strategies aim to deepen student engagement and encourage continued effort. Enhancing the game's storyline to promote emotional investment can significantly motivate students to progress and overcome challenges, with educational content appropriately embedded throughout (Dickey, 2011). Assigning unique personalities to each non-player character (NPC) allows for diverse interactions, enabling students to develop deeper connections through communication. Transforming dialogues with in-game characters into interactive conversations—where choices affect outcomes—can increase engagement and enhance information retention (Christensen et al., 2018). These choices can offer immediate feedback, such as changes in the affection level or favorability rating of NPCs, leading to different game endings, or providing additional hints for problem-solving within the game.

Practitioners should also consider incorporating incentives that acknowledge both effort and achievement (Rahimi et al., 2021). For instance, students could earn titles or badges based on their performance, whether by correcting previous mistakes or by making optimal choices on their initial attempts. After gameplay, instructors could host activities where students share the titles or badges they have earned, fostering peer interaction and further enhancing motivation to engage deeply with the game content.

5.3.5. Provide Clear Guidance and Support

To enable students to navigate challenges more effectively, it is essential to provide clear guidance and support tools within the game environment (Oren et al., 2020). In addition to implementing tutorials before introducing new game mechanics or educational content, practitioners should provide readily accessible in-game tools to assist students when they encounter difficulties and to enable them to monitor their progress. Such tools might include quest progress trackers, chat log histories, game maps, options to replay previous tutorials and instructions, and reviews of collected achievements. These features empower students to self-assess and navigate challenges independently, enhancing their learning experience.

Furthermore, for games which are designed for classroom integration (such as MHS), the development of a real-time instructor dashboard is essential (Nieland et al., 2021). This dashboard would allow educators to monitor students' progress and provide timely, appropriate interventions during or after gameplay. Given that an average classroom comprises approximately 15 to 20 students—based on observations during MHS field tests—it is impractical for instructors to offer individualized support without such technological assistance. Therefore, incorporating an instructor dashboard is crucial for facilitating effective teaching and ensuring that students receive the guidance they need within the DGBL environment.

5.3.6. Embed Adaptive Learning Experience

The oscillatory patterns observed in the ALE plots indicate that individual students may perceive game tasks differently regarding the tasks' complexity, learning difficulty, and cognitive load. To address these individual differences, implementing an adaptive system that utilizes real-time data to monitor fluctuations in student performance can provide timely support and adjust the game as needed.

For example, if a student consistently makes correct choices over several attempts, the task complexity can be increased to maintain an appropriate level of challenge and sustain engagement. Conversely, if a student struggles to make correct choices after a certain number of attempts, the task complexity can be decreased, alternative easier game mechanics can be introduced, or tutorials can be replayed. This adaptive approach allows students to learn and master the targeted knowledge at their own pace, ensuring that the game remains accessible and engaging for learners with varying abilities (Streicher and Smeddinck, 2016).

5.4. Limitations and Future Research

While our proposed analytic pipeline has proven effective and yielded several insightful findings, certain limitations that offer avenues for future research have been identified.

Firstly, our study utilized only one variable to represent students' argumentation skills, limiting our analysis to the overall outcome of this skill without providing insights into its distinct components. Future research should disaggregate the argumentation skill into multiple components and develop SAs for each segment. This approach would enable more precise formative feedback and a deeper understanding of student performance in different argumentation aspects.

Secondly, enhancements in our feature engineering process could further improve model performance. This study used factor analysis to generate features representing the interactive effects among various behavior types. However, this technique primarily captures linear interactions, leaving nonlinear interactions underexplored. Future research could incorporate advanced techniques such as autoencoders, kernel PCA, and t-SNE, which are better suited for generating nonlinear interactive representations. Additionally, our process of merging datasets from different behavior types involved normalizing and standardizing each dataset, followed by a merge based on student names. Exploring alternative data fusion methods could yield superior results when combining disparate datasets.

Regarding missing data, our current approach involved inserting zeros for null values, which may not be the most effective method. Future studies could employ more sophisticated imputation techniques, such as Hot-Deck, K-Nearest Neighbors (KNN), and Regression imputation, to enhance model training by better addressing missing data. Moreover, our feature set lacks elements representing student disengagement, off-task behavior, and guessing behaviors, potentially weakening the model's ability to identify low-performing learners who may require additional assistance. Future research should focus on generating features that specifically target the characteristics of low-level learners, which could significantly enhance the design of instructor dashboards, particularly in aiding teachers in identifying students who need additional support.

Another limitation of our study is that the proposed analytic pipeline's effectiveness was assessed within a single DGBL environment. Future research should expand this examination to multiple DGBL environments, particularly those featuring rich interaction mechanics and ill-structured problem-solving tasks, like MHS. Such environments offer diverse challenges and require nuanced solutions, making them ideal for assessing the pipeline's generalizability and robustness.

While we referenced several theories, such as cognitive load theory, to help interpret our ALE plot results, these theories did not explicitly guide our analytics. Future studies could adopt more theory-driven approaches, such as those adopted by Huang (2011) and Martinez-Garza and Clark (2017), focusing on specific theories like cognitive load theory to examine student learning in greater detail. For instance, future research could identify features from raw game logs to measure cognitive load, analyze its variation across different tasks and learners, and investigate its impact on learning outcomes. Such insights could guide the design of learning experiences that effectively manage cognitive load, potentially leading to improved learning outcomes.

Additionally, our analysis revealed that not all learning progress values were included in the final feature set used to train the model. This suggests that some learning progress values did not significantly influence the targeted learning objectives, or their impact was less substantial than that of certain behavior features, leading to their exclusion during the feature selection process. Ideally, all or most embedded scores would be included in the final model after feature selection. We may need to revisit the standards used to generate learning progress features and modify the formula guiding the transformation from raw game logs to these features to achieve this. The final standards and formula should be determined through expert knowledge, judgment, and iterative testing rather than relying on a single reference as in this study.

Furthermore, while video games are valued for their ability to provide diverse and engaging interactions that foster active learning, research on specific game mechanics tied to engagement outcomes remains sparse. As suggested by Boyle et al. (2016), the link between engagement and learning requires further investigation. Future research should focus on designing and generating additional learning progress values that measure student engagement at different game stages. This effort aims to construct links between engagement and learning and evaluate engagement differences across various game mechanics.

Finally, we also acknowledge the potential limitations of using ALE plots to interpret black-box models, such as the hard-voting ensemble learning models employed in this study (Wadoux and Molnar, 2022; Apley and Zhu, 2020). While ALE plots are useful for providing insights into how different features representing learning progress influence learning outcomes, offering an entry point for stakeholders like educators and game designers to understand relationships that might otherwise remain opaque, they are not without their limitations. This is particularly true when dealing with complex, nonlinear interactions or features that exhibit inherently non-monotonic effects.

When ALE plots yield unexpected patterns, such as bell-shaped curves or oscillating trends, caution must be exercised before attributing these findings solely to game design or feature interactions (Wadoux and Molnar, 2022). Such patterns may indicate not only unique feature effects but also potential artifacts or overfitting within the model. In these instances, the interpretability of ALE plots could be misleading if they fail to capture the nuanced relationships between features and outcome variables, especially in the high-dimensional feature space often found in DGBL environments.

Moreover, ALE plots operate under the assumption of conditional independence between features, which, when violated, can lead to inaccurate representations of feature effects (Apley and Zhu, 2020). This is particularly relevant in DGBL contexts, where features are often highly correlated due to the complex interplay of learning behaviors, strategies, and game mechanics. In the current study, although unsupervised learning techniques were employed to reduce intercorrelations between features, these correlations could not be entirely eliminated, potentially contributing to the unexpected patterns observed in the ALE plots. As such, interpretations derived from these plots should be supplemented by domain knowledge or further statistical investigation.

In light of these concerns, it is crucial to treat ALE plots as one component of a broader interpretative toolkit rather than as a definitive source of insights. In future research, we plan to explore alternative interpretability methods, such as Shapley values or feature interaction analysis, which can theoretically provide a more comprehensive understanding of feature effects

(Lundberg and Lee, 2017; Hassija et al., 2024). By triangulating findings across multiple interpretative methods, we believe researchers and practitioners can more confidently distinguish true design flaws from limitations inherent in the interpretative approach.

6. Acknowledgments

This work is supported by the US Department of Education's Institute of Education Sciences (R305A150364) and the i3 program (U411C140081). The views expressed are those of the project team and do not reflect the funders. The authors wish to express their gratitude to Dr. Sean P. Goggins for his invaluable guidance and insightful feedback throughout the research and writing process. We also express our sincere thanks to Dr. Eric Wulff for his significant contributions to designing and refining the rubrics for generating learning progress features, which have greatly enriched this work. We are deeply appreciative of the team at *Adroit Studios Gaming Lab at the University of Missouri* for their support, resources, and facilities, as well as their expert feedback regarding the Mission HydroSci project. Their collaboration has been instrumental in the success of this research. Special thanks go to the editors of the *Journal of Educational Data Mining*, Dr. Agathe Merceron and Dr. Philip I. Pavlik Jr., and copyeditor Chelsea Porter for their meticulous editing and proofreading efforts, which have enhanced the clarity and quality of this paper. We are also grateful to Dr. Nigel Bosch for his assistance in resolving technical issues related to converting the manuscript to HTML format, which greatly facilitated the copyediting process. Finally, we thank Angela Jarecki for her creative contributions in designing and refining the pipeline diagram presented in this paper, which effectively illustrates the research process. The authors also acknowledge the anonymous reviewers for their constructive feedback and thoughtful suggestions, which have significantly improved the final manuscript.

7. Declaration of Generative AI Software tools in the writing process

During the preparation of this work, the author(s) used ChatGPT for grammatical clarification in the subsections "Machine Learning and Artificial Intelligence Application in Digitial Game-based Learning" and "Assessment Within Game-based Learning." After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

Abd-Alsabour, N. 2018. On the role of dimensionality reduction. *J. Comput. 13,* 5, 571–579.

Abdi, H. and Williams, L. J. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics 2,* 4, 433–459.

Afyouni, I., Murad, A., and Einea, A. 2020. Adaptive rehabilitation bots in serious games. *Sensors 20,* 24, 7037.

Akpan, B. 2020. Classical and operant conditioning—ivan pavlov; burrhus skinner. In *Science education in theory and practice: an introductory guide to learning theory*, B. Akpan, Ed. Springer, Cham, 71–84.

AKRAM, B., MIN, W., WIEBE, E., MOTT, B., BOYER, K. E., AND LESTER, J. 2018. Improving stealth assessment in game-based learning with lstm-based analytics. In *Proceedings of International Conference on Educational Data Mining*, M. Y. Kristy Elizabeth Boyer, Ed. International Educational Data Mining Society, Buffalo, NY, 208–218.

ALAM, A. 2021. Should robots replace teachers? mobilisation of ai and learning analytics in education. In *2021 International Conference on Advances in Computing, Communication, and Control (ICAC3)*. IEEE, IEEE, Mumbai, India, 1–12.

ALTMANN, A., TOLOŞI, L., SANDER, O., AND LENGAUER, T. 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics 26,* 10, 1340–1347.

ALUTHMAN, E. S. 2016. The effect of using automated essay evaluation on esl undergraduate students' writing skill. *International Journal of English Linguistics 6,* 5, 54–67.

ANNETTA, L. A., MURRAY, M. R., LAIRD, S. G., BOHR, S. C., AND PARK, J. C. 2006. Serious games: Incorporating video games in the classroom. *Educause quarterly 29,* 3, 16–22.

APLEY, D. W. AND ZHU, J. 2020. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82,* 4, 1059–1086.

ASSIRI, A. S., NAZIR, S., AND VELASTIN, S. A. 2020. Breast tumor classification using an ensemble machine learning method. *Journal of Imaging 6,* 6, 39.

AYESHA, S., HANIF, M. K., AND TALIB, R. 2020. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion 59*, 44–58.

BADA, S. O. AND OLUSEGUN, S. 2015. Constructivism learning theory: A paradigm for teaching and learning. *Journal of Research & Method in Education 5,* 6, 66–70.

BAHRAMI, M. R., BAHRAMI, B., BEHBOODI, F., AND POURRAFIE, S. 2023. Teaching the future: The vision of ai/chatgpt in education. In *Agents and Multi-Agent Systems: Technologies and Applications*, G. Jezic, J. Chen-Burger, R. S. M. Kusek, R. J. Howlett, and L. C. Jain, Eds. Vol. 1. Springer, Singapore, Rome, Italy, 393–402.

BARAB, S. A., GRESALFI, M., AND INGRAM-GOBLE, A. 2010. Transformational play: Using games to position person, content, and context. *Educational researcher 39,* 7, 525–536.

BARTHOLOMEW, D. J., KNOTT, M., AND MOUSTAKI, I. 2011. *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons, Online.

BAUER, G. R., MAHENDRAN, M., WALWYN, C., AND SHOKOOHI, M. 2022. Latent variable and clustering methods in intersectionality research: systematic review of methods applications. *Social psychiatry and psychiatric epidemiology 57*, 1–17.

BELLAND, B. R., WALKER, A. E., AND KIM, N. J. 2017. A bayesian network meta-analysis to synthesize the influence of contexts of scaffolding use on cognitive outcomes in stem education. *Review of educational research 87,* 6, 1042–1081.

BELLOTTI, F., KAPRALOS, B., LEE, K., MORENO-GER, P., AND BERTA, R. 2013. Assessment in and of serious games: An overview. In *Advances in Human-Computer Interaction*, A. B. Barreto, Ed. Vol. 2013. Hindawi Limited London, UK, United Kingdom, London, UK, United Kingdom, 1–11.

BIEDERMANN, D., CIORDAS-HERTEL, G.-P., WINTER, M., MORDEL, J., AND DRACHSLER, H. 2023. Contextualized logging of on-task and off-task behaviours during learning. *Journal of Learning Analytics 10,* 2, 115–125.

BIRD, K. A., CASTLEMAN, B. L., MABEL, Z., AND SONG, Y. 2021. Bringing transparency to predictive analytics: A systematic comparison of predictive modeling methods in higher education. *AERA Open 7*, 23328584211037630.

Blikstein, P. and Worsley, M. 2016. Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics 3,* 2, 220–238.

Block, J. H. and Burns, R. B. 1976. Mastery learning. *Review of research in education 4*, 3–49.

Botelho, A. F., Baker, R. S., and Heffernan, N. T. 2019. Machine-learned or expert-engineered features? exploring feature engineering methods in detectors of student behavior and affect. In *Proceedings of the Twelfth International Conference on Educational Data Mining*, C. Lynch and A. Merceron, Eds. International Educational Data Mining Society., Montréal, Canada, 1–4.

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C., and Pereira, J. 2016. An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education 94*, 178–192.

Braad, E., Žavcer, G., and Sandovar, A. 2016. Processes and models for serious game design and development. In *Entertainment Computing and Serious Games: International GI-Dagstuhl Seminar 15283, Dagstuhl Castle, Germany, July 5-10, 2015, Revised Selected Papers*, R. Dörner, S. Göbel, M. Kickmeier-Rust, M. Masuch, and K. Zweig, Eds. Springer, Springer Cham, Online, 92–118.

Breiman, L. 2001. Random forests. *Machine learning 45*, 5–32.

Caballero-Hernández, J. A., Palomo-Duarte, M., Dodero, J. M., and Gaševic, D. 2024. Supporting skill assessment in learning experiences based on serious games through process mining techniques. In *International Journal of Interactive Multimedia and Artificial Intelligence*, D. Verdú, Ed. Vol. 8. Universidad Internacional de La Rioja (UNIR) in Spain, Spain, 146–159.

Calixto, I., Rios, M., and Aziz, W. 2019. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Association for Computational Linguistics, Florence, Italy, 6392–6405.

Carbonaro, M., Cutumisu, M., Duff, H., Gillis, S., Onuczko, C., Schaeffer, J., Schumacher, A., Siegel, J., Szafron, D., and Waugh, K. 2006. Adapting a commercial role-playing game for educational computer game production. In *GameOn North America Conference*. Vol. 13. Institute for Operations Research and the Management Sciences (INFORMS), Monterey, CA, USA, 1–8.

Carvalho, M. B., Bellotti, F., Berta, R., De Gloria, A., Sedano, C. I., Hauge, J. B., Hu, J., and Rauterberg, M. 2015. An activity theory-based model for serious games analysis and conceptual design. *Computers & education 87*, 166–181.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., and Lopez, A. 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing 408*, 189–215.

Chaiklin, S. et al. 2003. The zone of proximal development in Vygotsky's analysis of learning and instruction. *Vygotsky's educational theory in cultural context 1,* 2, 39–64.

Champion, C. and Elkan, C. 2017. Visualizing the consequences of evidence in bayesian networks. *arXiv preprint arXiv:1707.00791 1,* 1, 1–9.

Chen, T. and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen, and R. Rastogi, Eds. Association for Computing Machinery, New York, NY, United States, San Francisco, USA, 785–794.

CHOPADE, P., EDWARDS, D., KHAN, S. M., ANDRADE, A., AND PU, S. 2019. Cpsx: using ai-machine learning for mapping human-human interaction and measurement of cps teamwork skills. In *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*. IEEE, IEEE, Arlington, VA, USA, 1–6.

CHRISTENSEN, D. S., JAKOBSEN, M., AND KRAUS, M. 2018. The engagement effect of players' agency over their characters' motivation. In *Interactivity, Game Creation, Design, Learning, and Innovation: 6th International Conference, ArtsIT 2017, and Second International Conference, DLI 2017, Heraklion, Crete, Greece, October 30–31, 2017, Proceedings 6*, N. V. Anthony L. Brooks, Eva Brooks, Ed. Springer, Springer, Crete, Greece, 184–193.

CIOLACU, M., TEHRANI, A. F., BINDER, L., AND SVASTA, P. M. 2018. Education 4.0-artificial intelligence assisted higher education: early recognition system with machine learning to support students' success. In *2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging(SIITME)*. IEEE, IEEE, Online, 23–30.

CLARK, D. B., TANNER-SMITH, E. E., AND KILLINGSWORTH, S. S. 2016. Digital games, design, and learning: A systematic review and meta-analysis. *Review of educational research 86,* 1, 79–122.

CONATI, C., PORAYSKA-POMSTA, K., AND MAVRIKIS, M. 2018. Ai in education needs interpretable machine learning: Lessons from open learner modelling. In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden*, A. W. Been Kim, Kush R. Varshney, Ed. arXiv, Stockholm, Sweden, 21–27.

COVITT, B. A., GUNCKEL, K. L., AND ANDERSON, C. W. 2009. Students' developing understanding of water in environmental systems. *The Journal of Environmental Education 40,* 3, 37–51.

DARVENKUMAR, T. AND DEVI, V. A. 2022. A review on the impact of using e-games for social development and language acquisition in the english classroom. *ECS Transactions 107,* 1, 15713.

DEWEY, J. 1974. *John Dewey on education: Selected writings*. University of Chicago, Chicago, Illinois, USA.

DICKEY, M. D. 2011. Murder on grimm isle: The impact of game narrative design in an educational game-based learning environment. *British Journal of Educational Technology 42,* 3, 456–469.

DIVJAK, B. AND TOMIĆ, D. 2011. The impact of game-based learning on the achievement of learning goals and motivation for learning mathematics-literature review. *Journal of information and organizational sciences 35,* 1, 15–30.

DOLECK, T., LEMAY, D. J., BASNET, R. B., AND BAZELAIS, P. 2020. Predictive analytics in education: a comparison of deep learning frameworks. *Education and Information Technologies 25*, 1951–1963.

DREY, T., JANSEN, P., FISCHBACH, F., FROMMEL, J., AND RUKZIO, E. 2020. Towards progress assessment for adaptive hints in educational virtual reality games. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, and K. Yatani, Eds. Association for Computing Machinery, New York, NY, United States, Honolulu, HI, USA, 1–9.

DYULICHEVA, Y. Y., KOSOVA, Y. A., AND UCHITEL, A. D. 2020. The augmented reality portal and hints usage for assisting individuals with autism spectrum disorder, anxiety and cognitive disorders. In *Proceedings of the 3rd International Workshop on Augmented Reality in Education*, A. E. K. Oleksandr Yu. Burov, Ed. CEUR Workshop Proceedings, CEUR, Kryvyi Rih, Ukraine, 251–262.

ELISEYEV, A. AND AKSENOVA, T. 2019. Personalized adaptive instruction design (paid) for brain–computer interface using reinforcement learning and deep learning: simulated data study. *Brain-Computer Interfaces 6,* 1-2, 36–48.

ESERYEL, D., GE, X., IFENTHALER, D., AND LAW, V. 2011. Dynamic modeling as a cognitive regulation scaffold for developing complex problem-solving skills in an educational massively multiplayer online game environment. *Journal of Educational Computing Research 45,* 3, 265–286.

ESERYEL, D., LAW, V., IFENTHALER, D., GE, X., AND MILLER, R. 2014. An investigation of the interrelationships between motivation, engagement, and complex problem solving in game-based learning. *Journal of Educational Technology & Society 17,* 1, 42–53.

ESTEVEZ, J., GARATE, G., GUEDE, J., AND GRAÑA, M. 2019. Using scratch to teach undergraduate students' skills on artificial intelligence. *arXiv preprint arXiv:1904.00296. 1,* 1, 1–6.

FROMMEL, J., ROGERS, K., BRICH, J., BESSERER, D., BRADATSCH, L., ORTINAU, I., SCHABENBERGER, R., RIEMER, V., SCHRADER, C., AND WEBER, M. 2015. Integrated questionnaires: Maintaining presence in game environments for self-reported data acquisition. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, A. L. Cox, P. Cairns, R. Bernhaupt, and L. Nacke, Eds. Association for Computing Machinery, New York, NY, United States, Seattle, Washington, USA, 359–368.

GAIKWAD, P. 2022. Sustainable learning: Implications from the law of the minimum. *Spicer Adventist University Research Articles Journal 1,* 1, 18–25.

GEE, J. P. 2003. What video games have to teach us about learning and literacy. *Computers in entertainment (CIE) 1,* 1, 20–20.

GEORGIADIS, K., VAN LANKVELD, G., BAHREINI, K., AND WESTERA, W. 2019. Learning analytics should analyse the learning: proposing a generic stealth assessment tool. In *2019 IEEE Conference on Games (CoG)*. IEEE, IEEE, London, United Kingdom, 1–8.

GEORGIADIS, K., VAN LANKVELD, G., BAHREINI, K., AND WESTERA, W. 2020. On the robustness of stealth assessment. *IEEE Transactions on Games 13,* 2, 180–192.

GÖBEL, S., DE CARVALHO RODRIGUES, A., MEHM, F., AND STEINMETZ, R. 2009. Narrative game-based learning objects for story-based digital educational games. *narrative 14*, 16.

GÖBEL, S. AND MEHM, F. 2013. Personalized, adaptive digital educational games using narrative game-based learning objects. In *Serious Games and Virtual Worlds in Education, Professional Development, and Healthcare*, W. B. Klaus Bredl, Ed. Informaiton Science Reference, Online, 74–84.

GREY, S., GREY, D., GORDON, N., AND PURDY, J. 2017. Using formal game design methods to embed learning outcomes into game mechanics and avoid emergent behaviour. *International journal of game-based learning (IJGBL) 7,* 3, 63–73.

GRIS, G. AND BENGTSON, C. 2021. Assessment measures in game-based learning research: A systematic review. *International Journal of Serious Games 8,* 1, 3–26.

GUNCKEL, K. L., COVITT, B. A., AND ANDERSON, C. W. 2009. Learning a secondary discourse: shifts from force-dynamic to model-based reasoning in understanding water in socioecological systems. In *Proceedings of Learning Progressions in Science (LeaPS) Conference, Iowa City, IA*, A. W. G. Alicia C. Alonzo, Ed. LeaPS Learning Progressions in Science, Iowa City, IA, USA, 1–19.

GUPTA, A., CARPENTER, D., MIN, W., ROWE, J. P., AZEVEDO, R., AND LESTER, J. C. 2021. Multimodal multi-task stealth assessment for reflection-enriched game-based learning. In *Proceedings of the First International Workshop on Multimodal Artificial Intelligence in Education (MAIED 2021) At the 22nd International Conference on Artificial Intelligence in Education (AIED 2021), Online, June 14th, 2021*, D. D. Mitri, R. M. Maldonado, O. C. Santos, J. Schneider, K. A. M. Sanusi, M. Cukurova, D. Spikol, I. Molenaar, M. N. Giannakos, R. Klemke, and R. Azevedo, Eds. CEUR Workshop Proceedings, vol. 2902. CEUR-WS.org, Online, 93–102.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., and Hussain, A. 2024. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation 16,* 1, 45–74.

Heine, C. 2020. Towards modeling visualization processes as dynamic bayesian networks. *IEEE Transactions on Visualization and Computer Graphics 27,* 2, 1000–1010.

Henderson, N., Acosta, H., Min, W., Mott, B., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., and Lester, J. 2022. Enhancing stealth assessment in game-based learning environments with generative zero-shot learning. In *Proceedings of the 15th International Conference on Educational Data Mining*, A. Mitrovic and N. Bosch, Eds. International Educational Data Mining Society, Durham, United Kingdom, 171–182.

Henderson, N., Kumaran, V., Min, W., Mott, B., Wu, Z., Boulden, D., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., et al. 2020. Enhancing student competency models for game-based learning with a hybrid stealth assessment framework. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, A. Rafferty and J. R. Whitehill, Eds. International Educational Data Mining Society, Online, 92–103.

Hooshyar, D., Ahmad, R. B., Yousefi, M., Fathi, M., Horng, S.-J., and Lim, H. 2016. Applying an online game-based formative assessment in a flowchart-based intelligent tutoring system for improving problem-solving skills. *Computers & Education 94*, 18–36.

Hooshyar, D., Malva, L., Yang, Y., Pedaste, M., Wang, M., and Lim, H. 2021. An adaptive educational computer game: Effects on students' knowledge and learning attitude in computational thinking. *Computers in Human Behavior 114*, 106575.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. 2013. *Applied logistic regression*. John Wiley & Sons, Online.

Huang, W.-H. 2011. Evaluating learners' motivational and cognitive processing in an online game-based learning environment. *Computers in Human Behavior 27,* 2, 694–704.

Huang, X., Wu, L., and Ye, Y. 2019. A review on dimensionality reduction techniques. *International Journal of Pattern Recognition and Artificial Intelligence 33,* 10, 1950017.

Hyvärinen, A. and Oja, E. 2000. Independent component analysis: algorithms and applications. *Neural networks 13,* 4-5, 411–430.

Ifenthaler, D. 2014. Toward automated computer-based visualization and assessment of team-based performance. *Journal of Educational Psychology 106,* 3, 651.

Ifenthaler, D., Eseryel, D., and Ge, X. 2012. *Assessment for game-based learning*. Springer, Online.

Ishak, S. A., Hasran, U. A., and Din, R. 2023. Media education through digital games: a review on design and factors influencing learning performance. *Education Sciences 13,* 2, 102.

Jackson, L. C., O'Mara, J., Moss, J., and Jackson, A. C. 2018. A critical review of the effectiveness of narrative-driven digital educational games. *International Journal of Game-Based Learning (IJGBL) 8,* 4, 32–49.

Johnson, C. I., Bailey, S. K., and Van Buskirk, W. L. 2017. Designing effective feedback messages in serious games and simulations: A research review. In *Instructional Techniques to Facilitate Learning and Motivation of Serious Games*, P. Wouters and H. van Oostendorp, Eds. Springer Cham, Switzerland, 119–140.

Kamel, H., Abdulah, D., and Al-Tuwaijari, J. M. 2019. Cancer classification using gaussian naive bayes algorithm. In *2019 International Engineering Conference (IEC)*. IEEE, IEEE, Erbil, Iraq, 165–170.

KANAL, V., BRADY, J., NAMBIAPPAN, H., KYRARINI, M., WYLIE, G., AND MAKEDON, F. 2020. Towards a serious game based human-robot framework for fatigue assessment. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, F. Makedon, Ed. Association for Computing Machinery, New York, NY, United States, Online, 1–6.

KHENISSI, M. A., ESSALMI, F., AND JEMNI, M. 2013. Toward the personalization of learning games according to learning styles. In *2013 International Conference on Electrical Engineering and Software Applications*. IEEE, IEEE, Hammamet, Tunisia, 1–6.

KIM, Y. AND ROSENHECK, L. 2020. Reimagining assessment through play: A case study of metarubric. In *Re-imagining University Assessment in a Digital World*, M. Bearman, P. Dawson, R. Ajjawi, J. Tai, and D. Boud, Eds. Vol. 7. Springer, Cham, Switzerland, 263–288.

KIM, Y. J. AND IFENTHALER, D. 2019. Game-based assessment: The past ten years and moving forward. In *Game-Based Assessment Revisited*, D. Ifenthaler and Y. J. Kim, Eds. Springer, Cham, Switzerland, 3–11.

KIRSCH, I., LYNN, S. J., VIGORITO, M., AND MILLER, R. R. 2004. The role of cognition in classical and operant conditioning. *Journal of clinical psychology 60,* 4, 369–392.

KLEMA, V. AND LAUB, A. 1980. The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control 25,* 2, 164–176.

KUZNETSOV, V. S., MOISEIENKO, M. V., MOISEIENKO, N. V., ROSTALNY, B. A., AND KIV, A. E. 2020. Using unity to teach game development. In *Proceedings of the 1st Symposium on Advances in Educational Technology (AET 2020)*. Vol. 2. Science and Technology Publications, Lda, Online, 506–515.

LAFFEY, J., GRIFFIN, J., SIGOLOFF, J., SADLER, T., GOGGINS, S., WOMACK, A., WULFF, E., AND LANDER, S. 2019. Mission hydrosci: Meeting learning standards through gameplay. In *Proceedings of The International Conference on Computer-Supported Collaborative Learning (CSCL)*, G. N. Kristine Lund, Ed. International Society of the Learning Sciences (ISLS), Lyon, France, 1017–1020.

LAFFEY, J. M., SADLER, T. D., GOGGINS, S. P., GRIFFIN, J., AND BABIUCH, R. N. 2019. Mission hydrosci: Distance learning through game-based 3d virtual learning environments. In *Virtual Reality in education: Breakthroughs in research and practice*, I. R. M. A. (IRMA), Ed. IGI Global, Lyon, France, 623–643.

LEE, D. D. AND SEUNG, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *nature 401,* 6755, 788–791.

LEONARDOU, A., RIGOU, M., AND GAROFALAKIS, J. 2020. Techniques to motivate learner improvement in game-based assessment. *information 11,* 4, 176.

LIN, H.-C. K., LIN, Y.-H., WANG, T.-H., SU, L.-K., AND HUANG, Y.-M. 2021. Effects of incorporating augmented reality into a board game for high school students' learning motivation and acceptance in health education. *Sustainability 13,* 6, 3333.

LIN, P.-H. AND CHEN, S.-Y. 2020. Design and evaluation of a deep learning recommendation based augmented reality system for teaching programming and computational thinking. *IEEE Access 8*, 45689–45699.

LIU, T. AND ISRAEL, M. 2022. Uncovering students' problem-solving processes in game-based learning environments. *Computers & Education 182*, 104462.

LOH, C. S., SHENG, Y., AND IFENTHALER, D. 2015. Serious games analytics: Theoretical framework. In *Serious games analytics: Methodologies for performance measurement, assessment, and improvement*, D. I. Christian Sebastian Loh, Yanyan Sheng, Ed. Springer Cham, Switzerland, 3–29.

LÓPEZ, C. AND TUCKER, C. 2018. Toward personalized adaptive gamification: a machine learning model for predicting performance. *IEEE transactions on Games 12,* 2, 155–168.

LU, W., GRIFFIN, J., SADLER, T. D., LAFFEY, J., AND GOGGINS, S. P. 2023. Serious game analytics by design: Feature generation and selection using game telemetry and game metrics–toward predictive model construction. *Journal of Learning Analytics 10,* 1, 168–188.

LU, Z., CHIU, M. M., CUI, Y., MAO, W., AND LEI, H. 2023. Effects of game-based learning on students' computational thinking: A meta-analysis. *Journal of Educational Computing Research 61,* 1, 235–256.

LUCKIN, R. AND CUKUROVA, M. 2019. Designing educational technologies in the age of ai: A learning sciences-driven approach. *British Journal of Educational Technology 50,* 6, 2824–2838.

LUNDBERG, S. AND LEE, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, and R. Fergus, Eds. Curran Associates Inc. 57 Morehouse LaneRed Hook, NY, United States, Long Beach, California, USA, 1–10.

MARTINEZ-GARZA, M. M. AND CLARK, D. B. 2017. Investigating epistemic stances in game play with data mining. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS) 9,* 3, 1–40.

MCGAGHIE, W. C. AND HARRIS, I. B. 2018. Learning theory foundations of simulation-based mastery learning. *Simulation in Healthcare 13,* 3S, S15–S20.

MILADINOVIC, I., SCHEFER-WENZL, S., AND BAJIC-KERNDL, K. 2023. Impact of a mobile serious game on learning outcomes for complex problem solving skills. In *2023 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, IEEE, Salmiya, Kuwait, 1–7.

MIN, W., FRANKOSKY, M. H., MOTT, B. W., ROWE, J. P., SMITH, A., WIEBE, E., BOYER, K. E., AND LESTER, J. C. 2019. Deepstealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies 13,* 2, 312–325.

MISLEVY, R. J., ALMOND, R. G., AND LUKAS, J. F. 2003. A brief introduction to evidence-centered design. *ETS Research Report Series 2003,* 1, i–29.

MOORE, G. R. AND SHUTE, V. J. 2017. Improving learning through stealth assessment of conscientiousness. In *Handbook on digital learning for K-12 schools*, T. H. Ann Marcus-Quinn, Ed. Springer, Online, 355–368.

MOURI, K., OKUBO, F., SHIMADA, A., AND OGATA, H. 2016. Bayesian network for predicting students' final grade using e-book logs in university education. In *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, IEEE, University of Macau, Macao, China, 85–89.

MOUSAVINASAB, E., ZARIFSANAIEY, N., R. NIAKAN KALHORI, S., RAKHSHAN, M., KEIKHA, L., AND GHAZI SAEEDI, M. 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments 29,* 1, 142–163.

MULWA, C., LAWLESS, S., SHARP, M., ARNEDILLO-SANCHEZ, I., AND WADE, V. 2010. Adaptive educational hypermedia systems in technology enhanced learning: a literature review. In *Proceedings of the 2010 ACM Conference on Information Technology Education*, M. Stinson, N. Cregger, K. Baker, and R. Friedman, Eds. Association for Computing Machinery, New York, NY, United States, University of Maryland, College Park, Maryland, USA, 73–84.

NATEKIN, A. AND KNOLL, A. 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics 7*, 21.

National Research Council. 2013. *Monitoring progress toward successful K-12 STEM education: A nation advancing?* National Academies Press.

Naul, E. and Liu, M. 2020. Why story matters: A review of narrative in serious games. *Journal of Educational Computing Research 58,* 3, 687–707.

Nguyen, H. A., Hou, X., Stamper, J., and McLaren, B. M. 2020. Moving beyond test scores: Analyzing the effectiveness of a digital learning game through learning analytics. In *Proceedings of the Thirteenth International Conference on Educational Data Mining (EDM 2020)*, A. Rafferty and J. R. Whitehill, Eds. International Educational Data Mining Society, Online, 487–495.

Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. 2015. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE 104,* 1, 11–33.

Nie, H., Xiao, H. M., and Shang, J. J. 2014. A critical analysis of the studies on fostering creativity through game-based learning. In *Hybrid Learning. Theory and Practice: 7th International Conference, ICHL 2014, Shanghai, China, August 8-10, 2014. Proceedings 7*, S. K. S. Cheung, J. Fong, L. F. Kwok, J. Zhang, and R. Kwan, Eds. Springer, Cham, Shanghai, China, 278–287.

Nieland, T., Fehrenbach, A., Marowsky, M., and Burfeind, M. 2021. The teacher-centered perspective on digital game-based learning: Quantitative and qualitative evaluation methods from diverse disciplines. In *Game-Based Learning across the Disciplines*, D. I. Carmela Aprea, Ed. Springer, Switzerland, 341–362.

Oren, M., Pedersen, S., and Butler-Purry, K. L. 2020. Teaching digital circuit design with a 3-d video game: The impact of using in-game tools on students' performance. *IEEE Transactions on Education 64,* 1, 24–31.

Osborne, S. P., Radnor, Z., and Nasi, G. 2013. A new theory for public service management? toward a (public) service-dominant approach. *The American Review of Public Administration 43,* 2, 135–158.

Peng, T., Luo, Y., and Liu, Y. 2022. Ai-based equipment optimization of the design on intelligent education curriculum system. In *Wireless Communications and Mobile Computing*, S. Rani, Ed. Vol. 2022. Hindawi, Baffins Lane, Chichester, West Sussex PO19 1UD, United Kingdom, 1–13.

Peterson, L. E. 2009. K-nearest neighbor. *Scholarpedia 4,* 2, 1883.

Pivec, M., Dziabenko, O., and Schinnerl, I. 2003. Aspects of game-based learning. In *Proceedings of the 3rd International Conference on Knowledge Management, Graz, Austria*, K. T. Hermann Maurer, Ed. Vol. 304. Springer Verlag Heidelberg, Graz, Austria, 216–225.

Plass, J. L., Homer, B. D., MacNamara, A., Ober, T., Rose, M. C., Pawar, S., Hovey, C. M., and Olsen, A. 2020. Emotional design for digital games for learning: The effect of expression, color, shape, and dimensionality on the affective quality of game characters. *Learning and instruction 70*, 101194.

Prensky, M. 2001. Digital natives, digital immigrants part 2: Do they really think differently? *On the Horizon 1,* 1, 1–9.

Rahimi, S., Shute, V., Kuba, R., Dai, C.-P., Yang, X., Smith, G., and Fernández, C. A. 2021. The use and effects of incentive systems on learning and performance in educational games. *Computers & Education 165*, 104135.

Reeves, T., Romine, W., Laffey, J., Sadler, T., and Goggins, S. 2020. Distance learning through game-based 3d virtual learning environments: Mission hydro science. evaluation report for mission hydrosci. *Grantee Submission 1,* 1, 1–31.

Romero, M., Usart, M., and Ott, M. 2015. Can serious games contribute to developing and sustaining 21st century skills? *Games and culture 10,* 2, 148–177.

Rosé, C. P., McLaughlin, E. A., Liu, R., and Koedinger, K. R. 2019. Explanatory learner models: Why machine learning (alone) is not the answer. *British Journal of Educational Technology 50,* 6, 2943–2958.

Rowe, E., Asbell-Clarke, J., and Baker, R. S. 2015. Serious games analytics to measure implicit science learning. In *Serious games analytics: Methodologies for performance measurement, assessment, and improvement*, C. S. Loh, Y. Sheng, and D. Ifenthaler, Eds. Springer, Switzerland, 343–360.

Rowe, E., Asbell-Clarke, J., Baker, R. S., Eagle, M., Hicks, A. G., Barnes, T. M., Brown, R. A., and Edwards, T. 2017. Assessing implicit science learning in digital games. *Computers in Human Behavior 76*, 617–630.

Rowe, J. P., Shores, L. R., Mott, B. W., and Lester, J. C. 2010. Individual differences in gameplay and learning: a narrative-centered learning perspective. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, Y. P. Ian Horswill, Ed. Association for Computing Machinery, New York, NY, United States, Worcester Polytechnic Institute (WPI) in Worcester, Massachusetts, USA, 171–178.

Sadler, T. D., Nguyen, H., and Lankford, D. 2017. Water systems understandings: a framework for designing instruction and considering what learners know about water. *Wiley Interdisciplinary Reviews: Water 4,* 1, e1178.

Sakulkueakulsuk, B., Witoon, S., Ngarmkajornwiwat, P., Pataranutaporn, P., Surareungchai, W., Pataranutaporn, P., and Subsoontorn, P. 2018. Kids making ai: Integrating machine learning, gamification, and social context in stem education. In *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE).* IEEE, IEEE, Wollongong, Australia, 1005–1010.

Saleh, A., Chen, Y., Hmelo-Silver, C., Glazewski, K., Mott, B., Taylor, R., Rowe, J., and Lester, J. 2019. Supporting collaborative problem solving in a game-based learning environment. In *The Proceedings of the International Conference on Computer-Supported Collaborative Learning (CSCL).* International Society of the Learning Sciences (ISLS), Lyon, France, 1029–1032.

Sanchez, D. R. and Lee, C. A. 2022. Understanding the challenges of game-based training: Recommendations for moving research forward in game-based learning. In *Handbook of Research on the Influence and Effectiveness of Gamification in Education*, A. C. Moreira, O. Bernardes, and V. Amorim, Eds. IGI Global, Online, 541–578.

Sasada, T., Liu, Z., Baba, T., Hatano, K., and Kimura, Y. 2020. A resampling method for imbalanced datasets considering noise and overlap. *Procedia Computer Science 176*, 420–429.

Schölkopf, B., Smola, A., and Müller, K.-R. 1997. Kernel principal component analysis. In *Proceedings of the International Conference on Artificial Neural Networks*, W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, Eds. Springer, Alghero, Sardinia, Italy, 583–588.

Serhan, B., Said, B., Cheniti, L., and El Khayat, G. 2019. Personalization in serious games for assessment. In *The 12th Annual International Conference of Education, Research and Innovation(ICERI2019) Proceedings*. IATED, IATED, Seville, Spain, 4845–4852.

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., and Fernández-Manjón, B. 2017. Applying standards to systematize learning analytics in serious games. *Computer Standards & Interfaces 50*, 116–123.

Shaffer, D. W. 2006. Epistemic frames for epistemic games. *Computers & education 46,* 3, 223–234.

SHARMA, K., PAPAMITSIOU, Z., AND GIANNAKOS, M. 2019. Building pipelines for educational data using ai and multimodal analytics: A "grey-box" approach. *British Journal of Educational Technology 50,* 6, 3004–3031.

SHUTE, V., KE, F., AND WANG, L. 2017. Assessment and adaptation in games. In *Instructional Techniques to Facilitate Learning and Motivation of Serious Games*, H. v. O. Pieter Wouters, Ed. Springer Cham, Switzerland, 59–78.

SHUTE, V., RAHIMI, S., SMITH, G., KE, F., ALMOND, R., DAI, C.-P., KUBA, R., LIU, Z., YANG, X., AND SUN, C. 2021. Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *Journal of Computer Assisted Learning 37,* 1, 127–141.

SHUTE, V. AND VENTURA, M. 2013. *Stealth assessment: Measuring and supporting learning in video games*. The MIT Press, Online.

SHUTE, V. J. 2011. Stealth assessment in computer-based games to support learning. *Computer games and instruction 55,* 2, 503–524.

SHUTE, V. J. AND KIM, Y. J. 2014. Formative and stealth assessment. In *Handbook of Research on Educational Communications and Technology*, J. M. Spector, M. D. Merrill, J. Elen, and M. J. Bishop, Eds. Springer New York, NY, New York, NY, 311–321.

SHUTE, V. J. AND RAHIMI, S. 2021. Stealth assessment of creativity in a physics video game. *Computers in Human Behavior 116*, 106647.

SHUTE, V. J., VENTURA, M., BAUER, M., AND ZAPATA-RIVERA, D. 2009. Melding the power of serious games and embedded assessment to monitor and foster learning. *Serious games: Mechanisms and effects 2*, 295–321.

SHUTE, V. J., WANG, L., GREIFF, S., ZHAO, W., AND MOORE, G. 2016. Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior 63*, 106–117.

SLAVIN, R. E. 1987. Mastery learning reconsidered. *Review of educational research 57,* 2, 175–213.

SMITH, G., SHUTE, V., AND MUENZENBERGER, A. 2019. Designing and validating a stealth assessment for calculus competencies. *Journal of Applied Testing Technology 20,* S1, 52–59.

SOFLANO, M., CONNOLLY, T. M., AND HAINEY, T. 2015. An application of adaptive games-based learning based on learning style to teach sql. *Computers & Education 86*, 192–211.

SPRONCK, P., PONSEN, M., SPRINKHUIZEN-KUYPER, I., AND POSTMA, E. 2006. Adaptive game ai with dynamic scripting. *Machine Learning 63*, 217–248.

STADDON, J. E. AND CERUTTI, D. T. 2003. Operant conditioning. *Annual review of psychology 54,* 1, 115–144.

STEINEMANN, A. 2003. Implementing sustainable development through problem-based learning: Pedagogy and practice. *Journal of Professional Issues in Engineering Education and Practice 129,* 4, 216–224.

STEINMAURER, A., SACKL, M., AND GÜTL, C. 2021. Engagement in in-game questionnaires-perspectives from users and experts. In *2021 7th International Conference of the Immersive Learning Research Network (iLRN)*. IEEE, IEEE, University of Westminster, London, UK, 1–7.

STREICHER, A. AND SMEDDINCK, J. D. 2016. Personalized and adaptive serious games. In *Entertainment Computing and Serious Games: International GI-Dagstuhl Seminar 15283, Dagstuhl Castle, Germany, July 5-10, 2015, Revised Selected Papers*, R. Dörner, S. Göbel, M. Kickmeier-Rust, M. Masuch, and K. Zweig, Eds. Springer Cham, Springer Cham, Dagstuhl Castle, Germany, 332–377.

SUN, C.-T., WANG, D.-Y., AND CHAN, H.-L. 2011. How digital scaffolds in games direct problem-solving behaviors. *Computers & Education 57,* 3, 2118–2125.

SUNARYA, P. A. 2022. Machine learning and artificial intelligence as educational games. *International Transactions on Artificial Intelligence 1,* 1, 129–138.

SWELLER, J. 1994. Cognitive load theory, learning difficulty, and instructional design. In *Learning and Instruction*, Elsevier, Ed. Vol. 4. Elsevier, 295–312.

SWELLER, J. 2011. Cognitive load theory. In *Psychology of Learning and Motivation*. Vol. 55. Elsevier, Online, 37–76.

TADAYON, M. AND POTTIE, G. J. 2020. Predicting student performance in an educational game using a hidden markov model. *IEEE Transactions on Education 63,* 4, 299–304.

THARWAT, A. 2016. Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition 3,* 2, 145–180.

TUMENAYU, O. O., SHABALINA, O., KAMAEV, V., AND DAVTYAN, A. 2014. Using agent-based technologies to enhance learning in educational games. In *Proceedings of the International Conference e-Learning 2014. Multi Conference on Computer Science and Information Systems (Lisbon, Portugal, July 15-19, 2014)*, M. Baptista Nunes and M. McPherson, Eds. International Association for Development of the Information Society, Lisbon, Portugal, 149–155.

VERLEYSEN, M. AND FRANÇOIS, D. 2005. The curse of dimensionality in data mining and time series prediction. In *Computational Intelligence and Bioinspired Systems. 8th International Work-Conference on Artificial Neural Networks, IWANN 2005, Proceedings*, F. S. Joan Cabestany, Alberto Prieto, Ed. Springer Berlin, Heidelberg, Springer Berlin, Heidelberg, Vilanova i la Geltrú, Barcelona, Spain, 758–770.

WADOUX, A. M.-C. AND MOLNAR, C. 2022. Beyond prediction: methods for interpreting complex models of soil variation. *Geoderma 422*, 115953.

WANG, A., RAMASWAMY, V. V., AND RUSSAKOVSKY, O. 2022. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, United States, COEX exhibition center in Seoul, South Korea, 336–349.

WANG, L.-H., CHEN, B., HWANG, G.-J., GUAN, J.-Q., AND WANG, Y.-Q. 2022. Effects of digital game-based stem education on students' learning achievement: a meta-analysis. *International Journal of STEM Education 9,* 1, 1–13.

WANG, Y., YAO, H., AND ZHAO, S. 2016. Auto-encoder based dimensionality reduction. *Neurocomputing 184*, 232–242.

WATTENBERG, M., VIÉGAS, F., AND JOHNSON, I. 2016. How to use t-sne effectively. *Distill 1,* 10, e2.

WU, W.-H., HSIAO, H.-C., WU, P.-L., LIN, C.-H., AND HUANG, S.-H. 2012. Investigating the learning-theory foundations of game-based learning: a meta-analysis. *Journal of Computer Assisted Learning 28,* 3, 265–279.

WU, X., CHEN, J., YU, F., YAO, M., AND LUO, J. 2019. Joint learning of multiple latent domains and deep representations for domain adaptation. *IEEE transactions on cybernetics 51,* 5, 2676–2687.

YANG, C. C., CHEN, I. Y., AND OGATA, H. 2021. Toward precision education. *Educational Technology & Society 24,* 1, 152–163.

YANG, D., ZARGAR, E., ADAMS, A. M., DAY, S. L., AND CONNOR, C. M. 2021. Using interactive e-book user log variables to track reading processes and predict digital learning outcomes. *Assessment for Effective Intervention 46,* 4, 292–303.

YANG, K.-H. 2017. Learning behavior and achievement analysis of a digital game-based learning approach integrating mastery learning theory and different feedback models. *Interactive Learning Environments 25,* 2, 235–248.

YANG, Y.-T. C. 2012. Building virtual cities, inspiring intelligent citizens: Digital games for developing students' problem solving and learning motivation. *Computers & Education 59,* 2, 365–377.

YING, C., QI-GUANG, M., JIA-CHEN, L., AND LIN, G. 2013. Advance and prospects of adaboost algorithm. *Acta Automatica Sinica 39,* 6, 745–758.

ZEBARI, R., ABDULAZEEZ, A., ZEEBAREE, D., ZEBARI, D., AND SAEED, J. 2020. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends 1,* 1, 56–70.

ZHU, J. AND ONTAÑÓN, S. 2020. Player-centered ai for automatic game personalization: Open problems. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*, G. N. Yannakakis, A. Liapis, P. Kyburz, V. Volz, F. Khosmood, and P. Lopes, Eds. Association for Computing Machinery, New York, NY, United States, Online, 1–8.

A. APPENDIX A: THE DIGITAL GAME-BASED LEARNING ENVIRONMENT: MISSION HYDROSCI

Mission HydroSci is a first-person 3D narrative adventure designed to instruct middle school students in Water Science and Scientific Argumentation in response to the NGSS, which emphasizes a novel approach to science education, prioritizing student engagement with disciplinary core ideas, cross-cutting themes, and scientific practices. MHS is rooted in the pedagogical approach known as "transformational play" (Barab et al., 2010), which posits that students' learning is enhanced when they assume the role of a character who must utilize subject matter knowledge to make decisions and take actions within an educational game or simulation. Subsequently, students can apply what they have learned in the virtual world to solve real-world problems that resemble the scenarios they encountered during gameplay. Additionally, the design of MHS incorporates a learning progressions methodology for sequencing gameplay activities and content, drawing upon extensive knowledge regarding students' progress in learning about water systems (Covitt et al., 2009; Gunckel et al., 2009; Sadler et al., 2017) and scientific argumentation (Osborne et al., 2013).

We have integrated a logging system in conjunction with the game to assess students' learning outcomes, provide formative feedback, and work as a foundation of a data-driven dashboard for instructors to better understand students' performances during gameplay effectively and unobtrusively. To guide the design of the logging system, we have utilized two frameworks: the Activity Theory-based Model of Serious Games (ATMSG) (Carvalho et al., 2015) and Experience API (xAPI) (Serrano-Laguna et al., 2017). ATMSG can establish a connection between in-game activities and educational objectives, ensuring that learners can acquire the targeted knowledge or skills by completing in-game activities. In our case, we combined ATMSG with the theory of transformational play to design authentic activities or tasks that reflect real-world contexts and challenges in water science and scientific argumentation, enabling learners to address real-world issues in these fields by playing the game. ATMSG provides a framework for breaking down complex and dynamic systems, such as serious games, into components that lay the foundation for fine-grained learning analytics. After defining what content the logging system should capture from the game based on ATMSG, we incorporated xAPI to establish the data formats to track learners' interactions during the game and save them in the MongoDB server. xAPI standards provide a promising solution to automatically or semi-automatically realize intelligent DGBL environments that foster learners' learning through tailored scaffolding and context-sensitive feedback base don each learner's real-time game logs.

Based on the curriculum topics around water science and scientific argumentation, MHS contains six modules. Each module has a unique curriculum topic and game map with distinctive landscapes. Designed around the curriculum topic, the MHS team renders each module with different pedagogical arrangements and game mechanics or tasks. Additional hints and learning materials, such as animated posters describing the aliens' technology and the new planet's history, are scattered around the game world, encouraging learners to explore the virtual world while learning to improve their engagement and learning motives. On average, middle-school students will take around 10 hours to complete the MHS under the guidance of instructors in the classroom setting. The detailed information related to each module's design can be seen in Table7.

Table 7: Detailed description of the curriculum topic, pedagogical arrangements and corresponding game mechanics or tasks.

	Curriculum Topic	Curriculum Arrangement	Game Mechanics (main quests)
Unit 1	Tutorial unit	Brings the introduction to the components of scientific argumentation, including claim, reasoning, and evidence.	Talk to each key non-playable character (NPC).
			Guided by the AI ARF, players will learn to open and get familiar of each in-game tool through menus or hotkeys.
			Guided by the AI ARF, players will know how to navigate the game world, such as walking, running, and jumping guided by the AI ARF.
			Guided by Dr. Toppo (one of the NPCs), players will open and get familiar with the interface of the argumentation system to understand how to construct a complete scientific argument.
Unit 2	Watersheds	Content knowledge: Interpreting topographic maps	Find The Team: After crash-landing on a new planet, the main character (controlled by the player) must locate the rest of the crew. To accomplish this, they must interpret the topographic map and carefully observe their surroundings. However, there is no wayfinding assistance provided during this quest.

		Content knowledge: Watersheds	Collect samples from eastern and western waterfalls: Based on the conversations with NPCs, players need to find the positions of the eastern and western waterfalls. By investigating the samples of two waterfalls, players need to collect appropriate evidence describing the characteristics of each waterfall. In this way, they could deduce the conditions of each waterfall's watershed and prepare later scientific argumentation or debate with NPCs.
		Content knowledge: Relationship between topography and surface water	Argue which watershed is bigger: Dr. Toppo (one of the NPCs) will invite players to the argumentation system to construct a complete argument that makes sense with collected evidence from the waterfalls. The argumentation system mimics the solar system, where the claim works as the sun, and reason and evidence work as planets around the sun. The planets represent evidence position in the further interstellar orbit than the planets representing reasons. Players need to choose the correct claim, reason and evidence from available choices displayed in the left corner of the system.
			Jasper's proposal: Through a conversation with Jasper (Another NPC), you will debate with him to determine if his proposal about the new place is logical with the information the player collected from the environment today.

		Scientific argumentation: Practicing and getting more familiar with components of argumentation.	CREI system: To fix the system of the AI ARF, players will enter into a system called CREI to practice the definitions of three components of scientific argumentation. Players will see a screen showing different sentences, and they need to judge which component the sentence represents by throwing balls in the direction showing the correct component.
		Scientific argumentation: Learning to support claims with evidence.	Argue which watershed is bigger: Players need to choose correct evidence to support the pre-decided claim and reason.
Unit 3	Surface water	Content knowledge: Interpreting watershed representations.	Sam's supplies: Players will meet Samantha (NPC) at her garden base as she is just starting. To help her build up the garden, players need to transport supplies to Sam's Garden base through the river. Players must deliver 4 crates to the river stream to finish the quest. There are two river streams where players must investigate their water flows to decide which is the correct stream to transport. After players deliver each crate to a certain river stream, a dialogue will pop-up showing the feedback on whether the stream is correct.

			Collect pumps from the alien ruins: After finding the pollutant source, Sam told us she found a huge tree near an intersection of the river branches and doubted that some river branches were also polluted by the battery core. To ensure her thought, we need to enter into an alien ruin to collect pumps that allow us to plant Sam's seeds into the mini gardens along the river to test which branch was polluted. Players need to apply what they learned regarding water flows to unlock those pumps to solve puzzles within the alien ruin. The general format of the puzzle is to find and carry a cube from the surroundings, put it into the water channel,l and guide it to the destination by managing the water flow direction through a controlling panel.
		Content knowledge: Movement of dissolved materials in surface water.	Trace the source of the pollutant: After receiving the supplies, Sam found the river is polluted. She provides players with sensors that will light red when the river spot is polluted and green when it's clean. Players need to take advantage of the sensors and investigate the characteristics of the river, such as water flow direction, whether in a river branch or its surrounding environment, to find the source of the pollutant, which is a crashed battery core.

			Plant seeds: After getting the pumps, players can plant Sam's seeds into the garden along the river to trace how the dissolved pollutant materials spread along the river flow. Players need to observe the river conditions to judge which mini garden to plant to trace the flow direction of the dissolved pollutant materials accurately. Each time players plant the seed will trigger a dialogue showing Sam's feedback regarding whether the mini-garden is polluted.
		Scientific argumentation: Learning how the reasoning works to connect the claim and evidence.	Convince Bill the pollutant is nearby: After finding the position of the battery core, Bill (NPC) will invite us to enter the argumentation system to construct a complete scientific argument to convince him where the battery core is. The players must choose the correct reasoning to connect the pre-decided evidence and claim logically within the system.

Unit 4	Ground water	Content knowledge: Groundwater Content knowledge: Soil types and permeability	Enter the ruins: Players meet Anderson (NPC) in a desert area of the new planet and are told we need to access the subsurface water. Players need to enter the alien ruin to activate 5 engine parts by solving puzzles to restore the power of a huge alien drill. The general format of the puzzle is that the players need to figure out the appropriate proportion of soil types, such as sandy, silt or clay, which creates the perfect condition to reserve groundwater and operate the panel to point to the correct proportion so that the players can control the water to flow into or out of the tank. This way, a potable cube can reach the tank's ground or float up from the bottom of the tank to trigger the switch for recharging the engine and unlocking the next room in the alien ruin.
		Content knowledge: Water table	Drill to the water table: After restoring the power of the huge drill, players need to apply the knowledge learned from solving previous puzzles to operate the control panel controlling the drill to go deep down into the geological layer of the ground where the water table exists. Dialogues from Anderson will give feedback on whether the players choose the correct ground layer.
			Fountains: After returning to Anderson's base, players meet Anderson and see three blueprints of the fountain design from his presentation. By carefully investigating each blueprint, players need to figure out which one is constructing a fountain that can pump groundwater from the ground.

		Content knowledge: Infiltration	Investigate the flood in the military base: After successfully pumping out the groundwater using the huge drill, Anderson finds a flood in the military base and doubts it's the players' responsibility. Players arrive at the military base to investigate the reason behind the flood and start a comprehensive exploration. After the exploration, players will notice that the armory and the warehouse of the military base are both seriously flooded. Then, the players must search carefully in those two rooms to collect as much information as possible to determine what triggered the flood.
			Who flooded the warehouse: Anderson starts a debate with us to figure out who should be responsible for the warehouse flood. Players will respond to his debate by entering the argumentation system to construct a complete scientific argument with the evidence collected.
			Who flooded the armory: Anderson starts a debate with us to figure out who should be responsible for the flood of the armory. Players will respond to his debate by entering the argumentation system to construct a complete scientific argument with the evidence collected.
		Scientific argumentation: Construct a complete scientific argumentation to defense players' thoughts.	Who flooded the warehouse: This time, players will construct a complete argument without any pre-decided components.
			Who flooded the armory: This time, players will construct a complete argument without any pre-decided components.

Unit 5	Water cycle	Content knowledge: Condensation Content knowledge: Evaporation	Explore the alien ruin in the tropical island: Players met Bill (NPC) on a tropical island of the new planet. Bill wants to set up a factory generating bottled water but is still determining how to make it. He asks the players to explore the alien ruin for related techniques. Players need to solve puzzles within the alien ruin to get the technology. The general format of the puzzles is that the players need to operate the controlling panel to decide if they should condense into or evaporate out from the water tank to guide a portable cube to a certain spot for unlocking the next room. In the final room of the alien ruin, players can get the technology they need.
			Set up the condensation and evaporation machines in Bill's factory: After mastering the knowledge of water condensation, evaporation, and precipitation, players can help Bill set up the bottling factory. In this quest, players must set up the correct machine (condensation or evaporation machine) on each factory floor according to Bill's arrangement and information collected from each factory floor. After deciding on each floor's machine, players will see a pop-up dialogue showing whether the player set up the correct machine. Theoretically, the factory can start bottling water production when each floor has the correct machine.

		Content knowledge: Precipitation	Precipitate salt from the water collected from the cave: After getting out of the alien ruin, players and Bill come into a cave where they start arguing about whether the water here contains salt. Players set up the equipment collected from the cave to precipitate salt from the water to prove the thought.
			Convince Bill: Players need to construct a scientific argument to convince Bill further to clarify why there is salt in the water.
		Scientific argumentation: Learn how to provide a counterargument to a faulty claim.	Convince Bill: In this argumentation time, Bill will start a complete scientific argument. Players must provide a counterargument to convince Bill that his claim is wrong according to collected facts (evidence).
Unit 6 (Still in developing at the second field test)	Water in engineered systems and summarization	This unit is the culminating experience for players. To solve the problems of this unit, players need to review information collected from the whole previous game process and knowledge learned until the current time point.	Players wake up to a declining planet caused by natural disasters resulting from a system-wide imbalance. The crew realizes that restarting the alien structures transferred the planet's water to an alien shuttle orbiting it as a moon. The players travel to the shuttle, where they must use their knowledge of the ancient alien culture, water systems, and energy transfer to solve the final challenge. The cinematic ending of the game depends on their success in restoring balance to the planet and the mission. After the final challenge, players can freely explore the environments, finish side quests, and play in their customized base.

B. Appendix B: The Analytic Process and Results Regarding Pre- and Post-test Results

To investigate whether there is a significant score gain between pre- and post-assessment, we first conducted the Shapiro-Wilk test to check if the scores are normally distributed. Results indicated that none of the scores conformed to a normal distribution, necessitating the application of the non-parametric Wilcoxon test as an alternative to the paired t-test. Table 8 shows significant score enhancements from pre- to post-test were observed in units 2, 3, 4, and 5. Additionally, the increases in the aggregate score for all items in the WSA and the cumulative score of the AA were statistically significant.

Table 8: Statements of pre-and-post test score differences. We used ”.”, ”*,” ”**,” and ”***” markers represent the significance levels when the p-value equals to 0.1, 0.05, 0.01, and 0.001, respectively, tested by the Wilcoxon test. No marker following the number within the “Mean difference” row means the difference between pre and post-test scores is not statistically significant.

	Unit 2	Unit 3	Unit 4	Unit 5	Sum content score	Argument score
Question number	6	3	4	10	23	12
Mean pre score	2.98 (1.39)	1.45 (0.98)	2.28 (1.03)	6.81 (2.26)	13.52 (4)	6.75 (2.43)
Mean post score	3.66 (1.46)	1.88 (1.01)	2.62 (1.04)	7.34 (2.34)	15.5 (4.64)	7.7 (7.7)
Mean difference	0.68***	0.42***	0.34***	0.53***	1.97***	0.95***

C. Appendix C: The Full Scoring Rubric Table We Based on To Generate Learning Progress Features

Table 9: The full version of the scoring rubric table for generating learning progress features.

Game Context	Corresponding Quest	Embedded Score Name	Calculation Standards

Unit 2	Argue which watershed is bigger: In this quest, students will enter into a 2-D system where they will generate a complete argumentation with three components - Evidence, Reasoning and Claim – by dragging and dropping available choices.	bigger-Arg-Score	2 Points: correct answer within 3 attempts;
			1 Points: correct answer within 4 attempts;
			0 Points: no correct answer or correct answer after more than 4 attempts.

	CREI system: In this quest, students will enter a new game area where they are asked to deliver or kick soccer ball into different directions. Each direction represents a component of a complete argumentation. Students need to make the right decision based on the information they got from dialogues with an in-game NPC.	CREIScore	1 Point: for each correct soccer ball delivery, students will get one score for this quest; -1/3 Points: for each incorrect soccer ball delivery, students will lose 1/3 point.
	Jasper's proposal: In this quest, students will have a conversion with one of the NPCs, named Jasper. Within this conversation, students need to accept or deny Jasper's claim by making choices along with appropriate reasons.	Jasper-Critique-Score	1 Point: for selecting "you forgot evidence," students will get 1 point; 0 Points: for either "Jasper you are right; or "Jasper you forgot the claim," students will get no score.

	Find the team: In this quest, students need to find the correct location where the team is gathering based on cluses got from conversations and the in-game topographic map.	find-Team-Ave-Score	2 Point: opening the map during completing the quest; 1 Point: finding the correct location of the team in 3 minutes or less; 0 points: for anything else.
Unit 3	Sam's supplies: In this quest, students need to deliver supplies spread on the riverbanks. Since the weight of those supplies, students find the most efficient way to deliver those supplies is to through them into the river and let them flow to the destination – Sam's lab. Students need to choose the correct river branch based on their knowledge regarding waterflow.	crate-Delivery-Score	1 Point: for correct crate placement; 0 Point: for incorrect crate placement.
	Plant seeds: Sam constructed nurseries along the riverbanks. Students are asked to plant seeds in the nurseries where the river branches, they close to are polluted by a pollutant source on the upstream of the river.	plantScore	1 Point: Selecting a correct pump location; -1/2 Points: selecting an incorrect pump location.
	Convince Bill the pollutant is nearby: In this quest, students will enter the 2-D argumentation system to construct a complete scientific argumentation to convince one of the NPCs, Bill, regarding where the pollutant source is.	up-stream-Arg-Score	2 Points: for getting correct argument within 3 tries; 1 Point: for getting correct argument within 6 tries; 0 Points: no correct argument or getting correct one with more than 6 tries.

Unit 4	Drill to the water table: Students are asked to solve a puzzle intend to choose the correct geographical layer of the ground for extracting water.	drill-Room-Score	1 Point: Select correct depth; 0 Points: for incorrect depth.
	Who flooded the armory: Students will enter the argumentation system to determine who caused the armory's flood by constructing scientific argumentation.	flood-Armory-Score	2 Points: for correct argument within 3 tries; 1 Point: for correct argument within 6 tries; 0 Points: for no correct argument or correct argument with more than 6 tries.
	Who flooded the warehouse: Students will enter the argumentation system to determine who caused the warehouse's flood by constructing another scientific argumentation.	ware-House-Score	1 Point: for selecting "your argument uses the wrong piece of evidence;" 0 Points: for selecting anything else.
	Fountains: In this quest, students need to choose the correct design plan for constructing fountains, which can pump groundwater successfully.	fountain-Score	1 Point: for selecting Fountain 1; 0 Points: for selecting anything else.
Unit 5	Convince Bill: in this quest, students need to build a scientific argumentation to explain to Bill the reasons for salt's presence in the water.	convince-Bill-Score	2 Points: for correct argument within 3 tries; 1 Point: for correct argument within 6 tries; 0 Points: for no correct argument or correct argument with more than 6 tries.

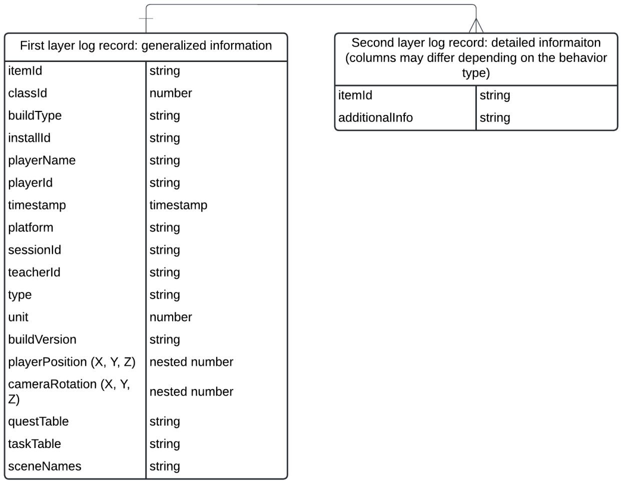
	Set up the condensation and evaporation machines in Bill's factory (First floor): In this quest, students need to help one of the NPCs – Bill – determine which machine is needed to be installed within each floor of Bill's water factory based on their knowledge regarding water cycle.	first-SetUp-Score	1 Point: for Select evaporator/ water to gas; 0 Points: for selecting anything else.
	Second floor	second-SetUp-Score	1 Point: for selecting condenser/ gas to water; 0 Points: for selecting anything else.
	Third floor	third-SetUp-Score	1 Point: for selecting evaporator/ water to gas; 0 Points for anything else.
	Fourth floor	forth-SetUp-Score	1 Point: for select condenser/ gas to water; 0 Points: for selecting anything else.

D. Appendix D: Supplementary Illustration Material for Methodology

D.1. Additional Illustrations about the Embedded Logging System

Figure 6 illustrates the basic data structure of the logging system. Additionally, Figure 7 provides example snippets from the raw log dataset, showcasing two different behavior types: movement and trigger (interaction with in-game objects).

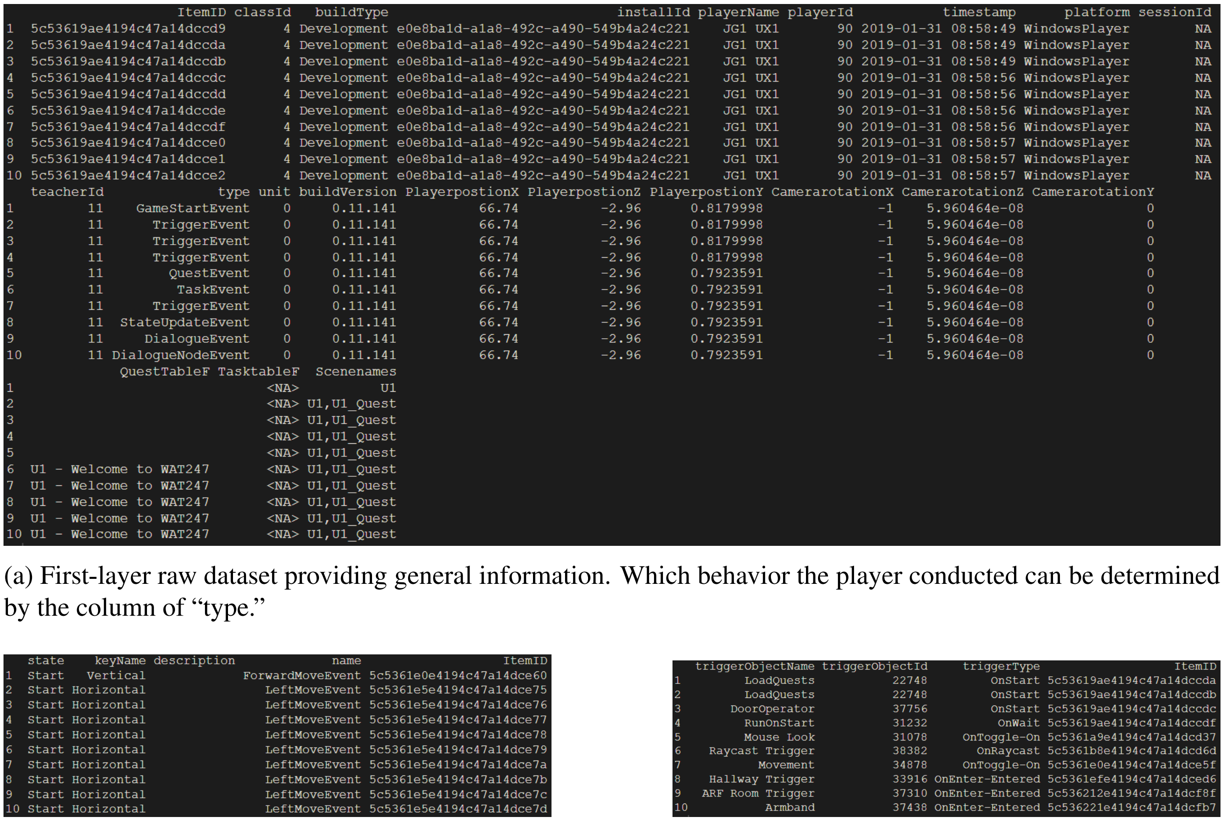
D.2. Additional Illustrations about the Final Observable or Raw Dataset



First layer log record: generalized information	
itemId	string
classId	number
buildType	string
installId	string
playerName	string
playerId	string
timestamp	timestamp
platform	string
sessionId	string
teacherId	string
type	string
unit	number
buildVersion	string
playerPosition (X, Y, Z)	nested number
cameraRotation (X, Y, Z)	nested number
questTable	string
taskTable	string
sceneNames	string

Second layer log record: detailed informaiton (columns may differ depending on the behavior type)	
itemId	string
additionalInfo	string

Figure 6: A diagram as a brief illustration regarding the structure of the raw log data.



```
                     ItemID classId   buildType                            installId playerName playerId           timestamp      platform sessionId
1  5c53619ae4194c47a14dccd9       4 Development e0e8ba1d-a1a8-492c-a490-549b4a24c221    JG1 UX1       90 2019-01-31 08:58:49 WindowsPlayer        NA
2  5c53619ae4194c47a14dccda       4 Development e0e8ba1d-a1a8-492c-a490-549b4a24c221    JG1 UX1       90 2019-01-31 08:58:49 WindowsPlayer        NA
3  5c53619ae4194c47a14dccdb       4 Development e0e8ba1d-a1a8-492c-a490-549b4a24c221    JG1 UX1       90 2019-01-31 08:58:49 WindowsPlayer        NA
4  5c53619ae4194c47a14dccdc       4 Development e0e8ba1d-a1a8-492c-a490-549b4a24c221    JG1 UX1       90 2019-01-31 08:58:56 WindowsPlayer        NA
5  5c53619ae4194c47a14dccdd       4 Development e0e8ba1d-a1a8-492c-a490-549b4a24c221    JG1 UX1       90 2019-01-31 08:58:56 WindowsPlayer        NA
6  5c53619ae4194c47a14dccde       4 Development e0e8ba1d-a1a8-492c-a490-549b4a24c221    JG1 UX1       90 2019-01-31 08:58:56 WindowsPlayer        NA
7  5c53619ae4194c47a14dccdf       4 Development e0e8ba1d-a1a8-492c-a490-549b4a24c221    JG1 UX1       90 2019-01-31 08:58:56 WindowsPlayer        NA
8  5c53619ae4194c47a14dcce0       4 Development e0e8ba1d-a1a8-492c-a490-549b4a24c221    JG1 UX1       90 2019-01-31 08:58:57 WindowsPlayer        NA
9  5c53619ae4194c47a14dcce1       4 Development e0e8ba1d-a1a8-492c-a490-549b4a24c221    JG1 UX1       90 2019-01-31 08:58:57 WindowsPlayer        NA
10 5c53619ae4194c47a14dcce2       4 Development e0e8ba1d-a1a8-492c-a490-549b4a24c221    JG1 UX1       90 2019-01-31 08:58:57 WindowsPlayer        NA
   teacherId              type unit buildVersion PlayerpostionX PlayerpostionZ PlayerpostionY CamerarotationX CamerarotationZ CamerarotationY
1         11    GameStartEvent    0     0.11.141          66.74          -2.96      0.8179998              -1    5.960464e-08               0
2         11      TriggerEvent    0     0.11.141          66.74          -2.96      0.8179998              -1    5.960464e-08               0
3         11      TriggerEvent    0     0.11.141          66.74          -2.96      0.8179998              -1    5.960464e-08               0
4         11      TriggerEvent    0     0.11.141          66.74          -2.96      0.8179998              -1    5.960464e-08               0
5         11        QuestEvent    0     0.11.141          66.74          -2.96      0.7923591              -1    5.960464e-08               0
6         11         TaskEvent    0     0.11.141          66.74          -2.96      0.7923591              -1    5.960464e-08               0
7         11      TriggerEvent    0     0.11.141          66.74          -2.96      0.7923591              -1    5.960464e-08               0
8         11  StateUpdateEvent    0     0.11.141          66.74          -2.96      0.7923591              -1    5.960464e-08               0
9         11     DialogueEvent    0     0.11.141          66.74          -2.96      0.7923591              -1    5.960464e-08               0
10        11 DialogueNodeEvent    0     0.11.141          66.74          -2.96      0.7923591              -1    5.960464e-08               0
              QuestTableF TasktableF  Scenenames
1                               <NA>          U1
2                               <NA> U1,U1_Quest
3                               <NA> U1,U1_Quest
4                               <NA> U1,U1_Quest
5                               <NA> U1,U1_Quest
6  U1 - Welcome to WAT247       <NA> U1,U1_Quest
7  U1 - Welcome to WAT247       <NA> U1,U1_Quest
8  U1 - Welcome to WAT247       <NA> U1,U1_Quest
9  U1 - Welcome to WAT247       <NA> U1,U1_Quest
10 U1 - Welcome to WAT247       <NA> U1,U1_Quest
```

(a) First-layer raw dataset providing general information. Which behavior the player conducted can be determined by the column of "type."

```
   state    keyName description             name                   ItemID
1  Start   Vertical             ForwardMoveEvent 5c5361e0e4194c47a14dce60
2  Start Horizontal                LeftMoveEvent 5c5361e5e4194c47a14dce75
3  Start Horizontal                LeftMoveEvent 5c5361e5e4194c47a14dce76
4  Start Horizontal                LeftMoveEvent 5c5361e5e4194c47a14dce77
5  Start Horizontal                LeftMoveEvent 5c5361e5e4194c47a14dce78
6  Start Horizontal                LeftMoveEvent 5c5361e5e4194c47a14dce79
7  Start Horizontal                LeftMoveEvent 5c5361e5e4194c47a14dce7a
8  Start Horizontal                LeftMoveEvent 5c5361e5e4194c47a14dce7b
9  Start Horizontal                LeftMoveEvent 5c5361e5e4194c47a14dce7c
10 Start Horizontal                LeftMoveEvent 5c5361e5e4194c47a14dce7d
```

(b) Second-layer raw dataset specific to Movement behavior type, providing additional information regarding Movement behaviors.

```
   triggerObjectName triggerObjectId     triggerType                   ItemID
1         LoadQuests           22748         OnStart 5c53619ae4194c47a14dccda
2         LoadQuests           22748         OnStart 5c53619ae4194c47a14dccdb
3       DoorOperator           37756         OnStart 5c53619ae4194c47a14dccdc
4         RunOnStart           31232          OnWait 5c53619ae4194c47a14dccdf
5         Mouse Look           31078     OnToggle-On 5c5361a9e4194c47a14dcd37
6    Raycast Trigger           38382       OnRaycast 5c5361b8e4194c47a14dcd6d
7           Movement           34878     OnToggle-On 5c5361e0e4194c47a14dce5f
8    Hallway Trigger           33916 OnEnter-Entered 5c5361efe4194c47a14dced6
9   ARF Room Trigger           37310 OnEnter-Entered 5c536212e4194c47a14dcf8f
10           Armband           37438 OnEnter-Entered 5c536221e4194c47a14dcfb7
```

(c) Second-layer raw dataset specific to Trigger behavior type, providing additional information regarding Trigger behaviors.

Figure 7: Example snippets of the raw log dataset specific to two different behavior types-Movement and Trigger.

Table 10: The descriptions regarding what features are involved within the final observable or raw feature set for our current study. Notably, for the behavior type of "Task completion," the task can be replayed in one gameplay trial, so there is no interaction frequency dataset under this behavior type. Also, for the behavior type of "Hotkey usage," because pressing a hotkey is a one-time action, it contains limited information to calculate the duration of pressing a specific hotkey, which makes us decide not to include the interaction speed dataset under this behavior type. Additionally, for the behavior type of "Map exploration," since the linear quest design of the game, there's nearly no need for students to explore a specific map more than one time, making us not include the frequency dataset under this behavior type.

Behavior Type	**Sub-Datasets**	**Description**
Task completion	Interaction speed	Each variable under this feature set represents the time duration a student used to complete a specific task.
	Interaction share	The time duration used to complete a particular task is divided by the total duration used to complete a unit or the whole game.
Argumentation	Interaction frequency	Each column under this dataset represents how often a student conducted a specific action within the argumentation system. The action can be dragging or dropping a particular node of choice into one component (evidence, reasoning, or claim) of a complete argument, Hovering on a particular choice node for reading detailed information, opening a specific in-system tool for seeking help, and submitting an answer with a successful or failed outcome
	Interaction speed	What is a student's average time to read the information for a particular choice node or in-system tool?
	Interaction share	The total frequency for a particular node or in-system tool divides the total frequency of all available nodes or tools.
Hotkey usage	Interaction frequency	Each column represents the frequency a student pressed a specific hotkey to reach a particular function quickly.
	Interaction share	The frequency of a specific hotkey divided by the total counts of all hotkey usage.
Tool menu usage	Interaction frequency	Each column reflects the frequency of a certain tool a student referred to for checking information.
	Interaction speed	Each variable indicates the average time duration a student used a specific tool.
	Interaction share	The total time duration of a specific tool usage is divided by the total duration of using all tools.

Dialogue reading	Interaction frequency	Each column represented the frequency of a particular dialogue (identified by dialogue ID) when a student went through one game trial. Dialogues designed to show once (usually those driving story progress) are deleted from this dataset because there is no variance or new information to investigate within those dialogues' counts.
	Interaction speed	Each column reflects the average speed at which a student reads a particular dialogue.
	Interaction share	The total duration a student used to read a specific dialogue divided by the total duration a student used to read all dialogues within the unit or the whole game.
Item triggering	Interaction frequency	Each variable represents the frequency a student interacts with a particular in-game item.
	Interaction speed	Like hotkey usage, some item interactions are a one-time occurrence and happen quickly, so we decided not to involve those items' durations in this feature set. For items that can be interacted with for a certain amount of time, such as the supply crates needed to be delivered to a river, cubes worked as a key to solve puzzles in dungeon areas. The panel students must solve a puzzle to unlock a door; we included the interaction durations for those items (identified by the item ID) in this feature set.
	Interaction share	The total count of students interacting with a certain item (identified by item ID) divided by the total count of students interacting with all available items within a unit or the whole game.
Behavior type statement	Interaction frequency	Each column represents the frequency of a particular behavior type that happened during a student's one time of playing through a particular unit or the whole game.
	Interaction speed	Each column represents the duration of a particular behavior type that happened during a student's one time of playing through a particular unit or the whole game.
	Interaction share	Each variable reflects the percentage of one specific behavior type that happened compared to the total number of log records for a student's one time of gameplay of a particular unit or the whole game.

Map exploration	Interaction speed	i. Each variable indicates the percentage of explored game map size for a student compared to the total map size. ii. Each variable represents the time duration used to explore one particular map and divides the total time spent exploring all maps within one unit or the whole game (If the unit includes just one map, then this feature set will not be generated).
	Interaction share	Each column reflects the total time duration a student spent on a specific game map (identified by the scene name).

D.3. Additional Illustrations about the Involved Machine Learning Algorithms

Table 11: Selected feature extraction techniques and corresponding selection reasons.

Feature Extraction Techniques	Brief Reasons to Select
Principal Component Analysis (PCA) (Abdi and Williams, 2010)	PCA is a classic linear dimensionality reduction technique that can be used to reduce the number of features while retaining the most variance in the data. It can be used on frequency dataset and especially useful if many behaviors are correlated, allowing users to reduce the complexity without losing critical information. However, it assumes variables within the feature set are normally distributed and linearly correlated, which usually is hard to be satisfied when dealing with real-world dataset.
Singular Value Decomposition (SVD) (Klema and Laub, 1980)	SVD is a powerful linear algebra technique that can effectively reduce dimensionality for large datasets, keeping useful information, represented by several components, and eliminating noises. It can also effectively handle sparse data, such as zeros or very low values. Its variation – truncated SVD – allow for focusing on the most significant components without needing to compute the full decomposition, thus saving computational resources.
Independent Component Analysis (ICA) (Hyvärinen and Oja, 2000)	ICA is used to reveal hidden factors (components) that are statistically independent from each other. This method is especially useful if you suspect that the observed frequency of behaviors can be explained by independent latent factors.

Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999)	Because the feature sets we involved to represent human behaviors are non-negative, NMF is an appropriate method to consider. NMF decomposes the feature matrix into two lower-dimensional matrices with the constraint that all matrices have no negative elements. This can be particularly useful for part-based representation where each component can be interpreted as contributing to some parts of the data structure.
Kernal PCA (Schölkopf et al., 1997)	It extends the capabilities of standard PCA by using kernel methods to capture nonlinear relationships within the data. This allows Kernel PCA to uncover complex, higher-dimensional structures hidden in the frequencies, which might not be apparent when using linear methods, thus providing a more detailed and insightful representation of behavioral interactions and dependencies.
T-distributed Stochastic Neighbor Embedding (tSNE) (Wattenberg et al., 2016)	This technique is well-suited for visualization purposes and can also be useful for exploring the structure of data representing human behaviors in a reduced dimension space. T-SNE is particularly good at maintain local structure and can reveal clusters or groups in the data, which might correspond to distinct behavioral patterns.
U-map (McGaghie and Harris, 2018)	U-map is suitable for feature extraction in datasets containing human behaviors because it excels at preserving both local and global structures in high-dimensional data. This technique effectively maps complex patterns and relationships among behavioral frequencies into a lower-dimensional space, facilitating insights into clusters, continuities, and variations in behavior that are crucial for understanding underlying patterns and groupings in the data.
Autoencoders (Wang et al., 2016)	Using neural networks, Autoencoders is a powerful method for feature extraction. It is a type of neural network that learns to compress (encode) the data into a smaller representation and then decompress (decode) it back to the original form. The encoder part of the network could provide users with a new, potentially more meaningful set of features to use.

Table 12: Classifier algorithms joined the hard voting ensemble learning and brief illustration.

Classifier Algorithm	Involvement Reason
C-Support Vector Classification (SVC) (Cervantes et al., 2020)	Provides robustness in high-dimensional spaces and is capable of defining complex, nonlinear boundaries using kernel mechanisms.
Random Forest (Breiman, 2001)	Excellent for handling a mix of numerical and categorical data, provides robustness against overfitting, and doesn't assume data linearity.

Logistic Regression (Hosmer Jr et al., 2013)	Good baseline linear model for binary classification.
K-Nearest Neighbors (KNN) (Peterson, 2009)	Adds a non-parametric approach that can adapt well to the data's local structure.
Gaussian Naïve Bayes (Kamel et al., 2019)	Quick and effective, especially in high-dimensional spaces despite its assumption of feature independence.
XGBoost Classifier (Chen and Guestrin, 2016)	A gradient boosting framework that is highly efficient and effective, often outperforming other classifiers in structured datasets.
Gradient Boosting Classifier (Natekin and Knoll, 2013)	Similar to XGBoost but still distinct as it might handle certain types of data differently.
AdaBoost Classifier (Ying et al., 2013)	Focuses on increasing the weight of misclassified instances and often complements well with other types of errors made by other classifiers.
Linear Discriminant Analysis (LDA) (Tharwat, 2016)	Provides a good linear decision surface based on class separability, which is different from logistic regression.
Quadratic Discriminant Analysis (QDA) (Tharwat, 2016)	Useful when the decision boundary between classes is quadratic.

E. Appendix E: ALE Plots for All Selected Learning Progress Features

E.1. Analysis of ALE plots for Unit 2's Water Science Content Knowledge

The ALE plot in Figure 8a illustrates the relationship between posttest scores measuring students' water science knowledge and their performance in scientific argumentation within Unit 2. The plot reveals that both the lowest and highest score intervals negatively impact the likelihood of attaining a high-level class in the corresponding learning outcome. In contrast, an intermediate score range of -0.2 to 0.3 positively affects the probability of being in the high-level class. As indicated in the row (a) of Table 13, students who submit correct arguments within three attempts, but not on the first attempt, are more likely to achieve high-level outcomes as assessed by the posttest.

Figure 8b presents the ALE plot depicting the relationship between posttest scores and task performance within the CREi system. The plot and the corresponding illustration in row (b) of Table 13 show a dichotomous impact: extreme performance levels—either consistently incorrect or achieving a perfect score (directing 5 or 6 soccer balls correctly out of 6)—correlate positively with high-level learning outcomes, as indicated by values above the zero threshold. Conversely, intermediate performance is associated with a decreased probability of achieving high-level content knowledge, as reflected by ALE values below the zero threshold.

In Figure 8c, the ALE plot examines the relationship between posttest scores and students' performance in the task of finding the team's location. Along with row (c) of Table 13, it indicates a pronounced positive impact on posttest scores for the highest performance range, marked by ALE values above the zero threshold. Specifically, students who locate their team within three minutes and utilize the topographic map effectively are significantly more likely to attain a high-level understanding of Unit 2's water science content. In contrast, all other performance levels are associated with lower posttest outcomes, suggesting a greater likelihood of achieving only a low-level understanding of the material.

Table 13: Score intervals and corresponding in-game performances for all selected features represent students' learning progress.

Feature Name	Standardized score value interval	Students in-game performance
(a) BiggerArgScore (Scientific argumentation in Unit 2)	Leftmost interval	Disengage from this argumentation task without submitting any correct argument.
	Middle interval	A correct argument was submitted within three attempts.
	Rightmost interval	Submitting a correct argument at the first submission.
(b) CREIScore (Choosing the correct component to complete the scientific argumentation)	Leftmost interval	The student delivered all balls into wrong directions, indicating an incorrect selection of the argumentation component.

	Middle interval	The student demonstrated mixed accuracy in their direction selection, with some balls being delivered into the correct directions, while others were delivered into the incorrect directions.
	Rightmost interval	The students delivered all soccer balls into correct directions, which means the students always choosing the correct scientific argumentation component.
(c) findTeamAveScore (Finding the team location using topographic knowledge)	Leftmost interval	The student either abandoned the quest or located the team after exceeding three minutes without leveraging the topographic map for navigation.
	Middle interval	The student either located the team after more than three minutes but knowing how to use topographic map for navigation, or alternatively, managed to find the team within three minutes but did not employe the map to find the correct location.
	Rightmost interval	The student successfully located the team within a span of three minutes or less, concurrently exhibiting a competent understanding of the topographic map’s usage during the navigation process.
(d) PlantScore (Examine students’ knowledge regarding dissolvable material within water flow in Unit 3)	Leftmost interval	The student either failed to complete the quest or incorrectly installed all pumps in inappropriate locations.
	Middle interval	The student demonstrated partial accuracy by installing some pumps at the correct locations while incorrectly placing others.
	Rightmost interval	The student successfully installed all pumps at the correct locations, thereby demonstrating complete accuracy in the task.
(e) UpstreamArgScore (Scientific argumentation in Unit 3)	Leftmost interval	The student failed to complete the task or made more than six attempts without submitting a correct argument.

	Middle interval	The student successfully submitted a correct argument within the range of four to six attempts.
	Rightmost interval	Submitted a correct argument within 3 tries.
(f) crateDeliveryScore (Delivering crates to Sam's base based on water flow direction)	Leftmost interval	The student either failed to complete the quest or incorrectly installed all pumps in inappropriate locations.
	Middle interval	The student demonstrated partial accuracy by installing some pumps at the correct locations while incorrectly placing others.
	Rightmost interval	The student successfully installed all pumps at the correct locations, thereby demonstrating complete accuracy in the task.
(g) floodArmoryScore (2nd scientific argumentation in Unit 4)	Leftmost interval	The student failed to complete the task or made more than six attempts without submitting a correct argument.
	Middle interval	The student successfully submitted a correct argument within the range of four to six attempts.
	Rightmost interval	Submitted a correct argument within 3 tries.
(h) wareHouseScore (1st scientific argumentation in Unit 4)	Leftmost interval	The student failed to complete the task or made more than six attempts without submitting a correct argument.
	Middle interval	The student successfully submitted a correct argument within the range of four to six attempts.
	Rightmost interval	Submitted a correct argument within 3 tries
(i) sortSecondScore (Score for setting up the correct type of machine in unit 5 – The second one)	Leftmost interval	The student either failed to complete the quest or selected an inappropriate type of machine.
	Middle interval	After several iterations of the quest, the student eventually made an appropriate selection regarding the machine type to install.
	Rightmost interval	The student successfully chose the correct machine type during the initial attempt.

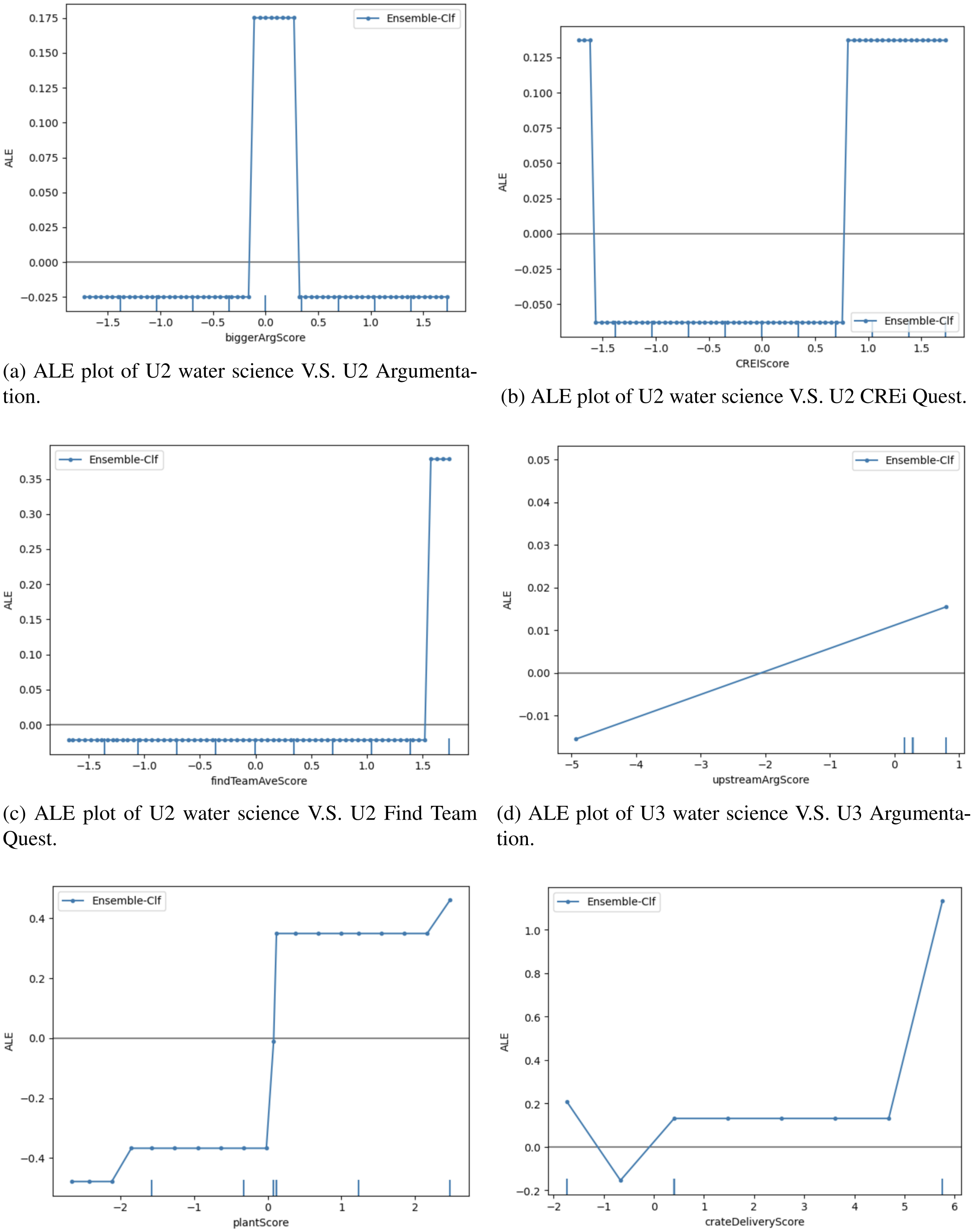
(j) thirdSetUpScore (Score for setting up the correct type) of machine in unit 5 – the third one	Leftmost interval	The student either failed to complete the quest or selected an inappropriate type of machine.
	Middle interval	After several iterations of the quest, the student eventually made an appropriate selection regarding the machine type to install.
	Rightmost interval	The student successfully chose the correct machine type during the initial attempt.
(k) FountainScore (Examine students' knowledge regarding soil types and ground water)	Leftmost interval	The student either failed to complete the quest or made an incorrect selection of the fountain design.
	Middle interval	After multiple attempts, the student successfully identified the correct fountain design.
	Rightmost interval	During the first attempt, the student accurately selected the correct fountain design.
(l) drillRoomScore (Score for drilling to the water table based on soil types)	Leftmost interval	The student either failed to complete the quest or was incorrectly drilled to an inappropriate water table depth.
	Middle interval	After multiple attempts, the student successfully identified the requisite fountain design.
	Rightmost interval	The student accurately selected the requisite fountain design at the first attempt.

E.2. Analysis of ALE plots for Unit 3's Water Science Content Knowledge

The analysis of ALE plots for Unit 3 reveals significant influences of the standardized scores "plantScore," "upstreamArgScore," and "crateDeliveryScore," on the posttest assessment of Unit 3's water science content knowledge.

The "plantScore" measures students' learning progress while installing four pumps for garden beds downstream of a large tree, based on their understanding of dissolvable materials in water. The ALE plot in Figure 8e shows fluctuating impacts on performance. Combined with row (d) of Table 13, we can observe that incorrect placement of more than two pumps decreases the likelihood of being classified as a high-level learner, whereas correctly installing more than two pumps significantly increases this probability. This is consistent with the description provided in the second row of Table 12.

For "upstreamArgScore," which assesses students' performance in a scientific argumentation task in Unit 3, there is a clear linear relationship between the standardized score and the



(a) ALE plot of U2 water science V.S. U2 Argumentation.

(b) ALE plot of U2 water science V.S. U2 CREi Quest.

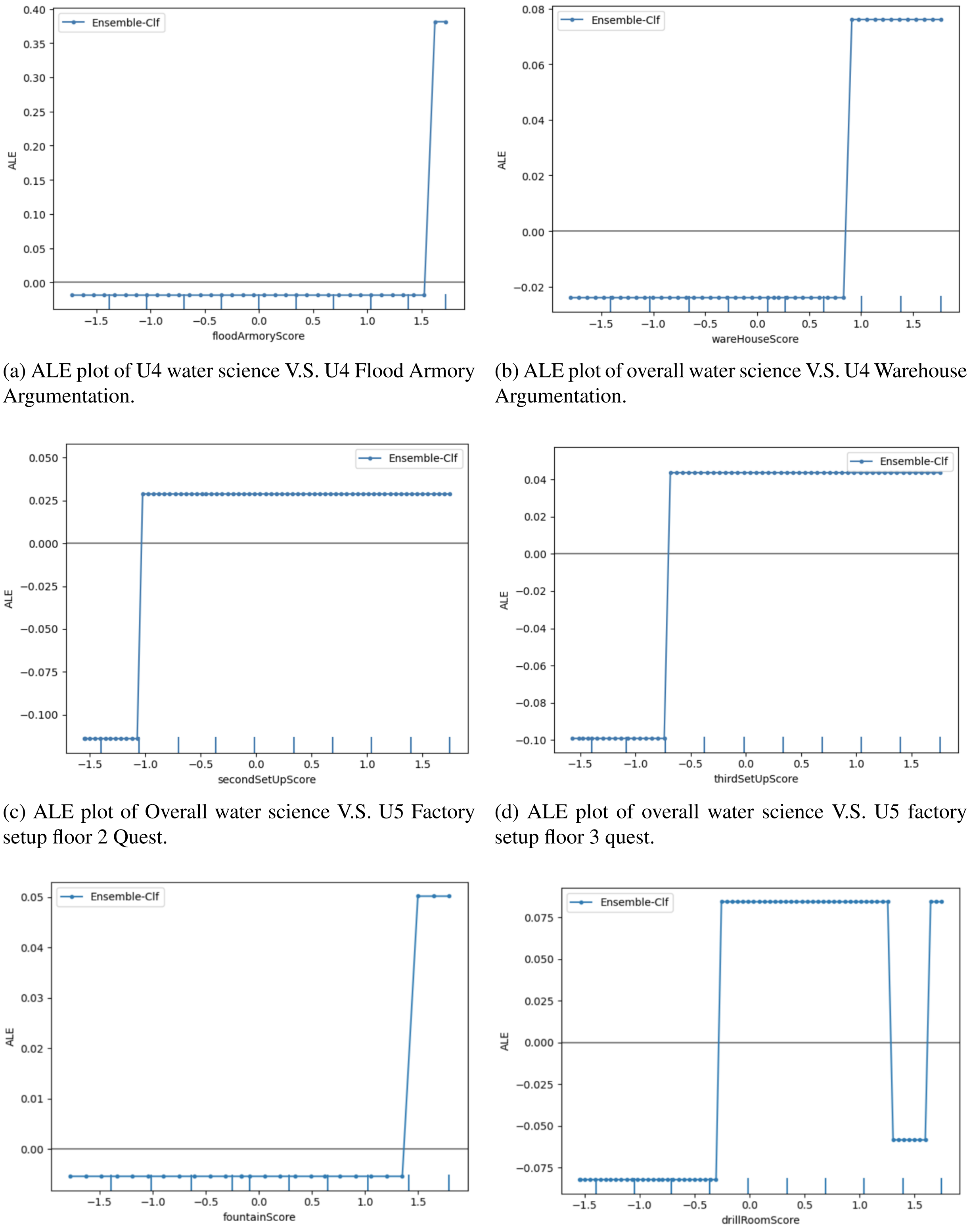
(c) ALE plot of U2 water science V.S. U2 Find Team Quest.

(d) ALE plot of U3 water science V.S. U3 Argumentation.

(e) ALE plot of U3 water science V.S. U3 Pump Installation Quest.

(f) ALE plot of U3 Water Science V.S. U3 Crate Delivery Quest.

Figure 8: Full selected ALE plots: part 1.



(a) ALE plot of U4 water science V.S. U4 Flood Armory Argumentation.

(b) ALE plot of overall water science V.S. U4 Warehouse Argumentation.

(c) ALE plot of Overall water science V.S. U5 Factory setup floor 2 Quest.

(d) ALE plot of overall water science V.S. U5 factory setup floor 3 quest.

(e) ALE plot of overall argumentation skill V.S. U4 fountain quest.

(f) ALE plot of overall argumentation skill V.S. U4 drill room quest.

Figure 9: Full selected ALE plots: part 2.

likelihood of being classified as a high-level learner, as depicted in Figure 8d. This task involves persuading a character, Bill, about the pollutant source based on evidence. According to row (e) of Table 13, students who submit the correct argument with fewer attempts are more likely to achieve high-level learner status in the posttest.

The "crateDeliveryScore" evaluates the accuracy of delivering crates based on the water flow in a task. Figure 8f and row (f) of Table 13 illustrate that delivering three crates incorrectly while one is correct has a notable negative impact on posttest scores. Conversely, correct delivery, particularly when all crates are delivered accurately, correlates positively with higher posttest scores. This pattern underscores the importance of accuracy in this task for better knowledge acquisition in Unit 3's water science content.

E.3. Analysis of ALE Plots for Unit 4's Water Science Content Knowledge

In Unit 4, the "floodArmoryScore" significantly influences the classification of students' learning outcomes in water science knowledge. This score measures students' progress in the second scientific argumentation task, where they must identify the individual responsible for triggering the armory flood. The ALE plot in Figure 9a shows that only the rightmost score interval positively impacts the probability of being classified as a high-level learner for Unit 4's water science knowledge. According to row (g) of Table 13, students who submit the correct argument within three attempts are more likely to be classified as high-level learners. In contrast, other score intervals, which include cases where students either did not submit a correct argument or took more than three attempts to do so, are associated with a decreased likelihood of achieving a high-level learning outcome in the posttest assessment.

E.4. Analysis of ALE Plots for Overall Water Science Content Knowledge

The analysis of ALE plots for overall water science knowledge reveals significant associations with specific learning progress measurements. The "wareHouseScore," which assesses students' ability to respond to an argument generated by an NPC to identify the person responsible for the warehouse flood, shows a notable correlation. As depicted in Figure 9b and discussed in row (h) of Table 13, students in the highest scoring interval—those who present the correct argument within three attempts—are more likely to be classified as high-level learners. In contrast, students who either exit the argumentation without submitting a correct argument or do so after more than three attempts have a reduced likelihood of being classified as high-level learners.

Additionally, the scores "secondSetUpScore" and "thirdSetUpScore," collected from tasks requiring students to choose between installing an evaporator or condenser in a factory, also demonstrate relevant trends, as shown by Figure 9c and Figure 9d. Based on row (i) and (j) of Table 13, both scores indicate a positive correlation between students' overall posttest water science knowledge and their performance in these tasks when they select the correct machine, regardless of the number of attempts. Conversely, incorrect machine selection is associated with lower overall posttest scores.

E.5. Analysis of ALE Plots for Overall Scientific Argumentation Skills

For overall scientific argumentation skills, the "fountainScore" has a significant impact on the classification probability of the corresponding learning outcome. As shown in Figure 9e and row (k) of Table 13, students in the highest score interval, who select the correct design on their first attempt, are more likely to be classified as high-level learners in argumentation skills. In contrast, students who do not choose correctly on their first attempt are less likely to achieve high-level outcomes in these skills.

The “drillRoomScore,” which measures students’ progress in selecting the appropriate ground level based on soil type for drilling groundwater, presents a complex relationship with overall scientific argumentation posttest scores. According to the ALE plot in Figure 9f and the last row of Table 13, success on the first attempt correlates with a higher likelihood of achieving high posttest scores in argumentation skills. Students who fail to find the correct ground level at all are less likely to attain high scores. Interestingly, the data shows that students who succeed on their second or third attempt have a lower probability of high posttest performance. However, those who succeed after more than three attempts, within a certain limit, exhibit a positive correlation with higher argumentation skill scores.