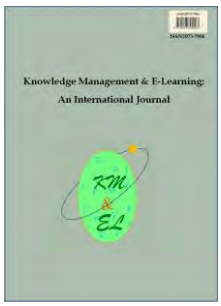
Knowledge Management & E-Learning, Vol.16, No.3. Sep 2024

Early prediction of mid-term and final scores using deep learning models

Danial Hooshyar
Tallinn University, Tallinn, Estonia
Nour El Mawas
Université de Lorraine, Metz, France
Yeongwook Yang
Gangneung-Wonju National University, Wonju, South Korea



Knowledge Management & E-Learning: An International Journal (KM&EL)
ISSN 2073-7904

Recommended citation:
Hooshyar, D., El Mawas, N., & Yang, Y. (2024). Early prediction of mid-term and final scores using deep learning models. *Knowledge Management & E-Learning, 16*(3), 398–423. https://doi.org/10.34105/j.kmel.2024.16.019

Knowledge Management & E-Learning, 16(3), 398–423

Early prediction of mid-term and final scores using deep learning models

Danial Hooshyar

School of Digital Technologies
Tallinn University, Tallinn, Estonia
E-mail: danial.hooshyar@tlu.ee

Nour El Mawas

Centre de Recherche sur les Médiations
Université de Lorraine, Metz, France
E-mail: nour.el-mawas@univ-lorraine.fr

Yeongwook Yang*

Department of Computer Science and Engineering
Gangneung-Wonju National University, Wonju, South Korea
E-mail: yeongwook.yang@gmail.com

*Corresponding author

Abstract: The use of learner modelling approaches is critical for providing adaptive support in educational computer games, with predictive learner modelling being among the key approaches. While adaptive supports have been shown to improve the effectiveness of educational games, improperly customized support can have negative effects on learning outcomes. To tackle these challenges, we present a novel approach, called DeepLM, that considers a series of time windows representing both sequences of learners' actions during gameplay and estimation of their current competencies (using stealth assessment) to model learners and accordingly predict their future performance. The approach employs a variant of deep neural networks to early predict learners' midterm and final scores simultaneously. The results show that using 20-50% of learners' action sequences can early predict their final scores, with a cross-validated convolutional neural network (CNN) achieving an area under the curve (AUC) and accuracy of 0.879 and 85.3%, respectively. The same model can also achieve high accuracy in predicting midterm and final scores at the same time, with an AUC and accuracy of 0.848 and 77.9%. Overall, the CNN model outperforms recurrent neural network, long short-term memory, and baseline multilayer perceptron (MLP) models in predicting learners' final performance and performs better than the baseline MLP model in predicting learners' midterm and final performance using both cross-validation and independent datasets. These findings show the potential of the proposed approach in accurately early predicting learners' performance, allowing educators and game designers to tailor interventions and support mechanisms that could lead to optimized learning outcomes.

Keywords: Stealth assessment; Predictive learner modelling; Deep neural

networks; Adaptive supports; Educational games

Biographical notes: Danial Hooshyar is an Associate Professor of Learning Analytics and Educational Data Mining at the Centre for Educational Technology, Tallinn University in Estonia. His research focuses on Artificial Intelligence for Education, with a particular emphasis on adaptive educational systems.

Nour EL Mawas is a Full Professor in Technology-enhanced Learning at the Centre de Recherche sur les Médiations, Université de Lorraine in France. received the Ph.D. degree in computer science from UTT, Troyes, in 2013. Her research interests include learning personalization, serious games, learning management systems, MOOCs, and lifelong learning.

Yeongwook Yang received a master's degree in computer science education, and a Ph.D. degree in the Department of Computer Science and Engineering from Korea University, Seoul, South Korea. He was a Research Professor at the Department of Computer Science and Engineering, Korea University, for one year. He was a Senior Researcher at the University of Tartu, Tartu, Estonia. He was an Assistant Professor at the Division of Computer Engineering, at Hanshin University. He is currently an Assistant Professor at the Department of Computer Science and Engineering, at Gangneung-Wonju National University. His research interests include information filtering, recommendation systems, educational data mining, and deep learning.

1. Introduction

Educational computer games have been shown to have the potential to improve learners' motivation, learning interests, knowledge gain, technology acceptance, etc. (e.g., Bakan et al., 2019; Blakely et al., 2009; El Mawas et al., 2020a; Hooshyar et al., 2021a; Ongoro & Mwangoka, 2022; Vlachopoulos & Makri, 2017). Because educational computer games (hereafter called educational games) can offer effective and engaging learning experiences, they have proven to be successful in different subject matters of different fields (e.g., Brezovszky et al., 2019; Hooshyar et al., 2021b; Pesare et al., 2016). Recently, there have been several research on equipping educational games with adaptivity that provide learners with support like offering personalized curricular sequencing, feedback, and hints (e.g., Liu et al., 2020; Malva et al., 2020). The cornerstone of delivering such adaptive supports is the learner model which sits behind the user interface of learning environments and – through analysing data on learner-system interactions – provides a representation of learners'

cognitive and non-cognitive characteristics. This model is then used to provide learners with an adaptive and optimal learning experience. Based on several research (e.g., Chen & Law, 2016; San Pedro et al., 2013), successful employment of adaptive supports requires accurate learner modelling and unsuitable adaptive supports can have a negative impact on the learning outcome of learners.

There are several learner modelling approaches when it comes to educational games (e.g., Hooshyar 2019a; Yannakakis & Togelius, 2018). For instance, knowledge tracing and stealth assessment both use learners' interaction with educational games to model their

knowledge and/or competencies. The stealth assessment approach (which usually employs Bayesian networks) considers an evidence-centred design framework and uses learners' data during gameplay to infer their knowledge. While effective, these approaches often require a direct mapping between knowledge content and learners' in-game actions (Kim et al., 2016). On the other hand, the knowledge tracing approach mostly uses Markov models to model learners' skill mastery (Anderson et al., 1995). This approach considers the probability of guessing/slipping to have a more realistic inference of learner knowledge. This allows educators and instructional designers to not only track skill proficiency but also gain insights into the factors influencing correct and incorrect responses. Nonetheless, such approaches mainly require historical data for parameter learning and may not provide an accurate prediction of learners' knowledge in situations where there are limited sequences of learners' actions.

Another prospective learner modelling approach is predictive learner modelling which uses learners' in-game data to infer their learner model and accordingly predict their future performance (Namoun & Alshanqiti, 2020). This approach does not require explicit incorporation of domain knowledge, which can be expensive in terms of human involvement. In recent years, there has been an increasing interest in using machine learning methods, including deep neural networks, for predictive learner modelling in educational games. Various studies, such as those conducted by Akram et al. (2018), Emerson et al. (2019), Henderson et al. (2020), Hooshyar et al. (2022), and Min et al. (2019), have explored the use of machine learning methods in this area. Among these methods, deep neural networks have demonstrated superior performance compared to other machine learning techniques for predicting learners' performance, as demonstrated by Hernández-Blanco et al. (2019).

Several studies have explored the use of deep neural networks and related techniques to predict learners' performance in educational games. For example, Min et al. (2019) utilized long short-term memory (LSTM) networks to capture temporal features of learners' gameplay data and predict their post-test performance. Geden et al. (2021) used natural language processing on learners' responses to reflection prompts within the game to develop a predictive model based on recurrent neural networks for post-test scores. Hooshyar et al. (2022) proposed an approach that integrates domain knowledge with deep neural networks to model learners' knowledge states during gameplay, taking into account their task completion success and algorithmic thinking strategies to predict their future performance. Another approach, presented by Lee-Cultura et al. (2020), utilizes multi-modal data to predict learners' academic performance early in an arithmetic operations game. These approaches leverage machine learning methods to automatically map learners' knowledge and actions in games and enable accurate prediction of their future performance.

Although predictive learner modelling approaches can alleviate challenges associated with manually relating learners' actions to content skills or knowledge, they may struggle to differentiate between a highly-scored solution resulting from mastery or chance. In educational games, a solution that receives a high score could be the result of a random or appropriate strategy, such as parallel thinking. An example of a random strategy that could be considered skill non-mastery is when a high-level solution is developed for a very simple task, e.g., unpurposefully travelling an empty path over and over. It should be considered that some random strategy-based solutions may result in obtaining scores and they do not always lead to losing scores.

To address such a challenge, one promising direction is to enrich predictive learner modelling with stealth assessment designed to model learners' current knowledge and

leverage it to early predict learners' performance. In other words, to consider the sequences of learners' actions as well as the estimation of their current knowledge (derived from Bayesian networks; see also Hooshyar et al., 2019b) in building a predictive learner modelling. This could leverage the current knowledge level of learners – which can logically be among the best indicators of their performance – in predicting their future performance. Moreover, coupling the sequence of estimation of learners' current knowledge with their in-game actions can relax the randomness challenge. The reason is that these estimations are inferred from carefully designed networks developed by subject matter experts linking knowledge content, gameplay strategies, game elements, learners' in-game actions, scores, etc.

In this study, a novel approach called DeepLM is proposed to predict learners' future performance in an educational game. The DeepLM employs a variant of deep neural networks to encode learners' knowledge states and competencies and differs from existing predictive learner modelling approaches in several ways. For instance, it predicts both midterm and final scores simultaneously at the early stages of the game by using limited sequences of learners' actions, and it experiments with different deep learning models, including convolutional networks and Multilayer perceptron, to find the best-performing one. This research contributes to the related works in several ways, including:

- Introducing a new approach called DeepLM that uses a limited sequence of learners' actions and their estimated competencies to predict both midterm and final scores simultaneously at the early stages of an educational game;
- Comparing the performance of different types of deep neural networks in predicting both single-step final scores and multi-step midterm and final scores;
- Testing the applicability of the validated models on separate datasets;
- Examining the impact of different lengths of action sequences on the prediction accuracy of the various deep neural network models.

2. Related works

2.1. Stealth assessment in educational games

Engagement plays a key role in learning and therefore good educational games should be engaging (Abdul Jabbar & Felicia, 2015; Parsons & Taylor, 2011; Xie et al., 2021). To provide and control engagement in educational games, it is vital to reliably measure learning during gameplay without disturbance of engagement. This information can later be used to support learning. Stealth assessment is one way to implant such reliable assessments into educational games (Shute & Kim, 2014; Shute & Ventura, 2013). Basically, using trace data of learners' interaction with educational games, stealth assessment infers learners' current knowledge according to an evidence-centred design framework or ECD (Kim et al., 2016; Mislevy et al., 2003). Overall, any assessment's main goal is to gather information that enables the evaluator to make a reliable and valid estimation of people's competencies (including knowledge, skills, abilities, and more). The ECD framework comprises computational and conceptual models working together that require identifying the assertions regarding competencies of learners, valid evidence of an assertation, and tasks or situations that obtain that evidence (Shute et al. 2016).

Stealth assessment meets the requirement of the ECD framework as it identifies certain learners' behavior during gameplay that can act as indicators of an assertation (evidence) and map them to the competency of learners (Shut & Ventura, 2013). Briefly, as learners interact with various tasks during gameplay, they leave a variety of digital traces or performance data that is automatically processed to induce learners' competency level. Learners' estimates of competency level can be used formatively and diagnostically for different purposes. Some examples include the adaptive and timely selection of non-player characters' behaviors or game level, and adaptive supports like individualized feedback, hints, and learning materials sequences. The dynamic nature of stealth assessment provides benefits like continual measurement of competencies of learners, task difficulty adjustment according to the performance of learners, and offering ongoing and timely feedback.

During the past few years, there have been several prototypes for the stealth assessment of learners in educational games. For example, Shute et al. (2010) and Shute and Kim (2012) proposed stealth assessment approaches based on the Bayesian network and investigated their feasibility by embedding them into educational games like Taiga Park and World of Goo, respectively. Besides, there are many educational games that follow the same line of work to provide adaptive support in educational games. For instance, the AutoThinking game is a computer game developed to promote lifelong learners' computational thinking. The game benefits from a Bayesian network decision-making algorithm that predicts learners' current competencies of computational thinking in real time and accordingly provides adaptive support in both learning and gameplay (Hooshyar et al., 2021a). While such a way of assessment eliminates the need for disrupting learning and has shown to be successful in practice, they are usually unable to use limited sequences of learners' actions in order to infer their knowledge in upcoming game tasks and accordingly predict their future performance.

2.2. Predictive learner modelling in educational games

Research has demonstrated promising potential for the development of predictive models that estimate learners' competencies and behaviors in educational games (Ha et al., 2012; Morshed Fahid et al., 2021; Wang et al., 2017). While both learner modelling and predictive learner modelling use currently available learners' data to infer their knowledge and skills, predictive learner modelling focuses on the prediction of learners' future performance rather than current skills and knowledge. Two of the most important learner modelling approaches are stealth assessment and knowledge tracing (Liu, 2022; Shute & Kim, 2014). As mentioned in the previous section, the stealth assessment approach considers an evidence-centred design framework and uses learners' data during gameplay to infer their knowledge, whereas the knowledge tracing approach mostly uses Markov models or sequence-based neural networks to model learners' mastery level of skills or knowledge components in adaptive learning systems (Corbett & Anderson, 1994).

When it comes to predictive learner modelling, in recent years, there has been growing literature on the early prediction of learner performance. Most of these works employ predictors like pre-test scores, survey data, and demographical data. For instance, Olivé et al. (2019) employed learners' data gathered up to a few days before assignment deadlines and accordingly predicted the timeliness of the submissions. In a different attempt, sequences of learners' actions were used by Jiménez et al. (2019) to early predict learners' dropout in a computer science program.

Lee-Cultura et al. (2020) developed a method that employs multi-modal data including Empatica E4 Wristband and eye tracker for predicting learners' academic performance in an educational game aimed at promoting arithmetic operations. They found that an ensemble learner could accurately predict learners' performance at an early stage. Geden et al. (2021) developed a predictive model to estimate learners' post-test scores in educational games, using natural language processing on learners' responses to in-game reflection prompts to enhance the predictive models. Their approach, which employed recurrent neural networks, achieved higher accuracy than other representations. Min et al. (2019) used LSTM networks to predict learners' post-test performance following interactions with an educational game. Their approach captured the temporal representation of learners' data and mapped learners' knowledge and actions in games using deep neural networks. The study demonstrated that their approach outperformed competitive baseline models in terms of early prediction capacity and accuracy. Hooshyar et al. (2022) proposed an approach that integrates domain knowledge with deep neural networks to model learners' knowledge states during gameplay. Their approach successfully predicted learners' performance early during gameplay using deep neural networks. Despite the success, Hooshyar et al. (2022)'s work face some challenges, such as: (1) merely utilizing learners' sequences of task IDs and their respective correctness to estimate learners' mastery of skills and predict their performance in upcoming game tasks; (2) employing fixed sequences of tasks and their correctness (e.g., 10, 15, and 19) and not being applicable to varying sequences; and (3) being uncapable to early predict learners' performances using limited sequences of their actions (e.g., using 20% of action sequences to early predict midterm and final scores).

The objective of our study is to expand the related works by leveraging diverse deep neural network models, enriched with stealth assessment, to anticipate learners' performance in educational games at an early stage. To this end, we propose a predictive learner modelling approach named DeepLM, which integrates in-game data along with current assessments of learners' knowledge and competencies to predict their performance. The approach includes the early prediction of learners' final scores in ongoing game episodes, as well as simultaneous multi-step prediction of midterm and final scores. Additionally, our approach experiments with different variants of deep neural networks (including convolutional networks) rather than simply selecting a sequence-based deep learning model like LSTM or RNN to find the best performing for the task.

3. DeepLM: Early prediction of learners' midterm and final scores

This section presents our proposed method for early prediction of learners' performance in educational games. The approach involves the segmentation of gameplay into multiple time windows that represent learners' sequential actions and estimation of their current competencies. To accomplish this, the DeepLM approach leverages a variant of deep neural networks to capture the latent knowledge states of learners and predict their midterm and final scores simultaneously. The subsequent section details the DeepLM approach in depth.

3.1. Early prediction task

In the AutoThinking game, learners are tasked with developing solutions to help a mouse evade cats and accumulate scores, with the potential to create up to 20 solutions of varying

quality during gameplay. Our study focuses on predicting learners' midterm and final scores using a limited sequence of their actions over time, specifically by utilizing the first n action sequences of learners who have completed the third level of the game (as the game only logs learner digital traces during the third level). To accomplish this task, we create time windows of game data for each learner, in which we extract feature vectors based on their solutions and actions at intervals determined by a maximum solution size τ and a percentage p of their action sequences. These time windows are used to generate collections of solutions and actions for each game episode, from which we can predict learners' midterm and final scores. Specifically, for single-step prediction or prediction of final scores, learners who obtained final scores lower than the mean were considered low performers and the rest as high performers. Therefore, according to whether a learner is a low or high performer, we assigned one ground truth label to each trajectory. Accordingly, 148 and 191 learners were labelled as low and high performers, respectively. For multi-step prediction or prediction of midterm and final scores at the same time, learners who obtained midterm and final scores lower than the mean were considered low performers and the rest high performers. This then allows us to assign two ground truth labels to each trajectory. Accordingly, 72 and 267 learners were labelled as low and high performers for midterm scores, while 148 and 191 learners were labelled as low and high performers for final scores.

3.2. The proposed approach

Fig. 1 shows the overall architecture of our proposed approach. In the first step, we gather and pre-process the sequence of solution submissions made by learners. We utilize a set of attributes associated with the final scores achieved by learners, including the number of collected objects, scores obtained for each action performed, and task and learner IDs, among others. As part of this process, we exclude game data from those learners who did not complete the game episode, as their final score is necessary to signify their final performance. Afterwards, for single-step prediction or prediction of the final score and multi-step prediction or prediction of both midterm and final scores simultaneously, we create varying time windows or sequences of input vectors using the sequence of learners' actions (including estimation of learners' current knowledge). Thereafter, we discretize learners' scores using their average scores and create low and high performers scores for every learner. We create labels for the single-step prediction using learners' final score category, and for the multi-step prediction, both midterm and final score categories are used. More specifically, we create time windows using 20%, 30%, 40%, 50%, 60%, and 70% of action sequences for each learner in the single-step prediction, while we develop 20%, 30%, 40%, and 50% of their action sequences for the multi-step prediction of midterm and final scores simultaneously. For learners' final scores, we consider their final solution number, and for their midterm score, we use both the final solution number and the size of their action sequences. For instance, for a learner with 10 developed solutions, using 20% of action sequences (in other words, two solution submissions), the score obtained at the end of solution number 10 is considered the final score, whereas the score gained at the end of solution six is considered the midterm score. Obviously, according to game conditions, objectives, and requirements, this way of determining midterm scores could be changed.

Upon creating nonstationary time windows, the subsequent step involves generating multiple datasets for the early prediction of a learner's final or midterm and final scores. The data is normalized using Z-transformation, followed by the

implementation of Nearest Neighbour imputation to predict and replace missing values, which primarily arise due to the varying nature of the created time windows. Once this is done, the data is split using shuffled sampling to create datasets for cross-validation and application (independent dataset, see section 4.2). Depending on the length of action sequences or time windows, the Synthetic Minority Oversampling Technique (SMOTE) is applied to balance the classes. SMOTE focuses on generating synthetic samples for the minority class by creating new instances that combine the features of existing minority class samples. This approach aids in creating a more balanced representation of the classes in the dataset, which is particularly important for training accurate machine learning models. By introducing these synthetic instances, SMOTE helps the model better understand and learn from the minority class, ultimately improving its ability to make accurate predictions.

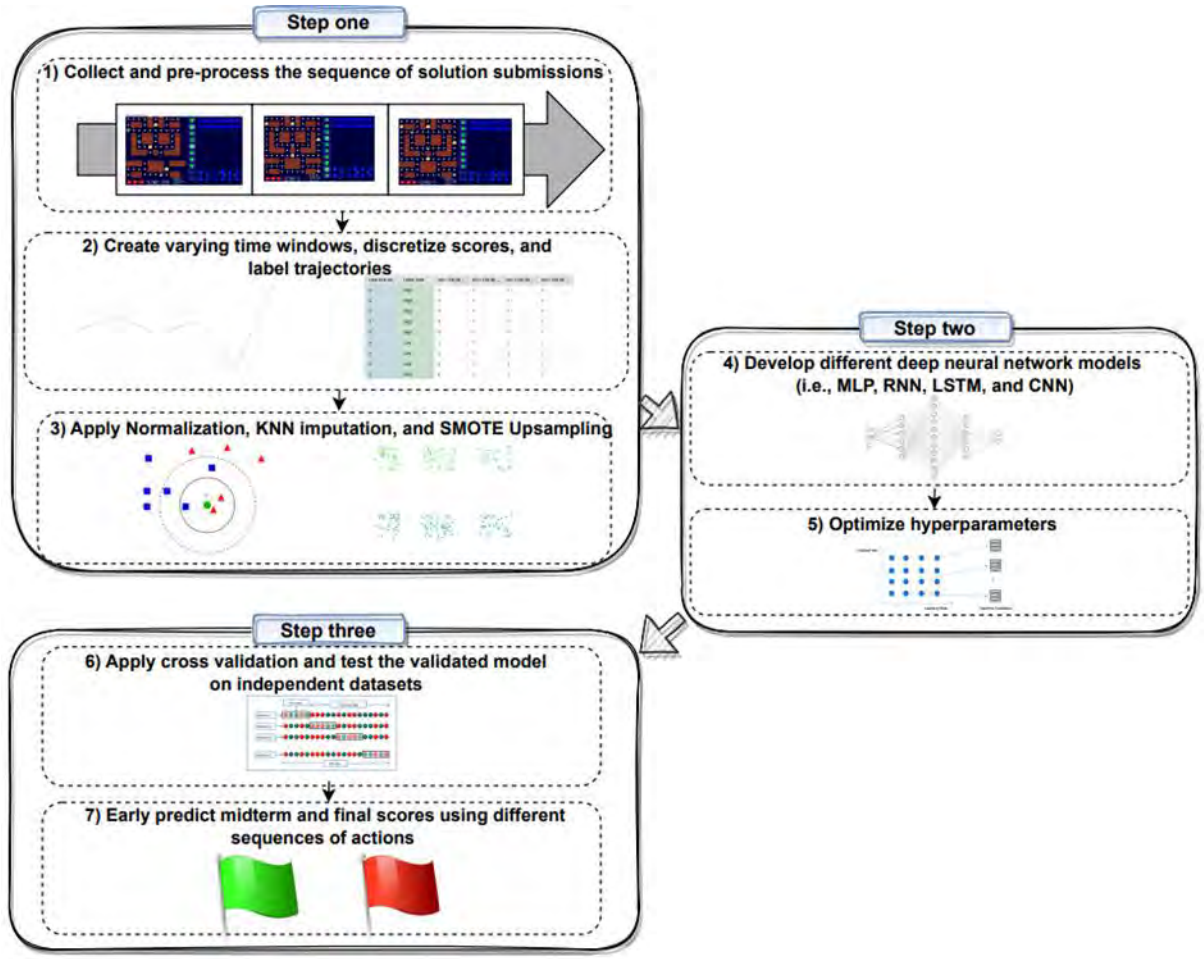


Fig. 1. The DeepLM approach for early prediction of midterm and final scores

In the second and third stages of our approach, we developed four deep learning models, including Multilayer perceptron (MLP or baseline model), Convolutional Neural Networks (CNN), Recurrent neural networks (RNN), and Long Short-Term Memory (LSTM) networks. We fine-tuned the models' parameters, such as the number of layers

and nodes, and learning rate, using an evolutionary approach based on a Genetic algorithm. We opted for evolutionary parameter optimization instead of grid and greedy search due to its advantage in cases where the ideal parameter ranges are uncertain. To evaluate the performance of our models, we employed a 10-fold cross-validation technique with shuffled sampling. Additionally, we assessed the practicality, stability, and robustness of the models using independent test sets.

4. Experimental evaluation

4.1. Context on the AutoThinking game for computational thinking

AutoThinking represents an educational game developed to enhance learners' prowess in computational thinking. It includes three levels, with the first two being mostly introductory and non-adaptive. Instead of relying on conventional programming languages, the game employs icons to symbolize programming concepts, thus removing the possibility of syntax errors. Additionally, AutoThinking provides adaptivity in both gameplay and the learning process by adaptively controlling the movements of Non-Playable Characters (NPCs) and offering feedback/hints, respectively. The game focuses on four important computational thinking skills: breaking problems into smaller steps (algorithmic thinking), making plans by noticing patterns, finding and fixing errors (debugging), and simulating solutions. It also teaches three basic concepts about programming: doing things step by step, making choices based on situations, and doing things over and over again (Hooshyar et al., 2019b).

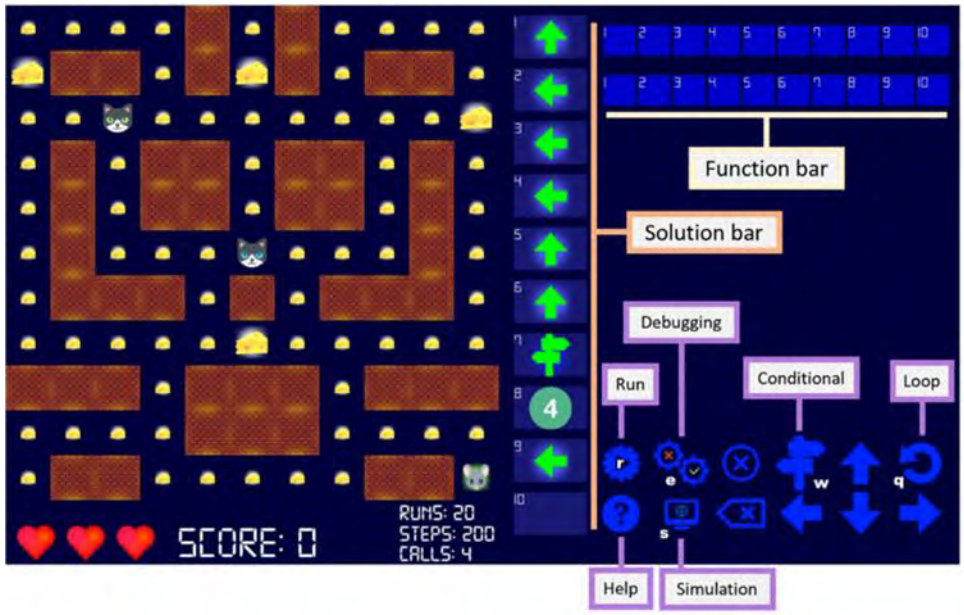


Fig. 2(a). A solution developed by a learner (taken from Hooshyar, 2022)

In the game, learners pretend to be a mouse and go through different levels. The goals are to collect cheese, get points, and avoid the NPCs (i.e., cats) in a maze. Learners can use up to 20 different strategies to get all 76 pieces of cheese. Solutions that incorporate

computational thinking concepts and skills, as well as navigating through non-empty tiles receive higher scores. Learners have the flexibility to develop various solutions, including using functions to save and apply patterns in different situations, and the game provides adaptive feedback and hints. Examples of a learner-developed solution, as well as video feedback generated by the game are illustrated in Fig. 2(a) and Fig. 2(b), respectively.

Fig. 2(b). Feedback generated for the situation (taken from Hooshyar, 2022)

4.2. Datasets

This study utilized log data collected from 427 Autothinking game learners from five countries (i.e., Estonia, France, South Korea, Taiwan, and South Africa) between December 2019 and April 2022, resulting in a dataset of 6199 solutions or examples. The learners varied in age and background, and some data were collected during experimental studies in France, Taiwan, and Estonia (e.g., El Mawas et al., 2020b; Hooshyar et al., 2021b), while others were collected from learners playing the game independently.

To evaluate the learners' computational thinking mastery, this study considered features related to their scores, such as the number of small and big cheese eaten, the frequency of simulation and debug usage, task IDs, arrow, loop, and conditional usage, command length (number of elements used in each solution), function bar usage, feedback and hints frequency, bumping into walls frequency, learner IDs, scores obtained for using computational thinking skills and concepts in a single solution, overall score for a single solution and game episode, and the Bayesian estimate of the learner's current knowledge or quality of the developed solutions and computational thinking concepts (for further information on the Bayesian estimates and other features, see also Hooshyar et al. 2019b).

After filtering out incomplete data, two datasets were created for cross-validation and independent application phases using shuffled sampling. The validation dataset contained 5272 solutions from 305 learners (134 low and 171 high performers before

upsampling, and 171 each after upsampling), while the independent application dataset contained 586 solutions from 34 learners (14 low and 20 high performers). Table 1 illustrates the basic statistics of the included attributes in datasets for cross-validation and independent application phases. To investigate the performance of the proposed approach in early score prediction using different sequences of learners' actions, different versions of each dataset (for validation and application phases) were developed, consisting of 20-70% of the learners' action sequences.

Table 1
Basic statistics of the datasets

Attributes	Validation			Application		
	Min	Max	Average	Min	Max	Average
Big cheese	0	2	.110	0	1	.094
Big cheese random	0	2	.10	0	2	.089
Small cheese	0	30	4.040	0	19	4.060
Simulation	0	1	.809	0	1	.814
Debug	0	1	.947	0	1	.947
Function	0	4	.030	0	2	.031
Command length	0	10	6.826	1	10	6.725
Task	0	20	6.329	0	19	4.060
Arrow	1	10	5.555	1	10	5.387
Loop	0	3	.419	0	3	.480
Conditional	0	3	.201	0	3	.174
Feedback	0	3	.183	0	3	.176
Hint	0	1	.099	0	1	.099
Bumping into walls	0	14	.455	0	10	.357
Knowledge estimates	0	1	.409	0	1	.418
Command score	0	490	31.150	0	360	33.480
Solution score	-990	1850	157.520	-930	1650	140.787
Final score	-1980	4033	1265.386	-1550	3929	1323.551

4.3. Experiment setting and evaluation

In this study, a computer with a single AMD Ryzen 5 PRO 4650U CPU and 16.0 GB memory was utilized to train the deep learning models. Stochastic Gradient Descent optimization using standard Backpropagation and ADAM updater with a learning rate of 0.01 and 10 epochs were employed to train the models. Regularization was also applied with L1 and L2 values set to 1. For the MLP (baseline) model, we used two fully connected layers with 50 neurons and the activation function of the Sigmoid. The RNN and LSTM models used recurrent and long-short term layers with 50 neurons and Sigmoid activation functions. The CNN model used a convolutional layer with 64 activation maps, Kernel size of 3, Stride size of 1, and a padding model of Truncated. On top of this, the model used a fully connected later with 10 neurons.

In order to assess the effectiveness of our models during the validation stage, we utilized *k*-fold cross-validation. This method is highly regarded for validating predictive models, particularly when the dataset is relatively small. We chose a value of $k = 10$ based on empirical evidence that suggests it produces test error rate estimates that do not suffer

from either excessive variance or high bias. In the application phase, we evaluated the cross-validated models using the independent game dataset. Ultimately, we compared the performance of various models using six distinct sequences of learners' actions.

5. Results and analysis

Four metrics of the area under the curve (AUC), accuracy, precision, and recall are used to evaluate the performance of the models. The AUC is the measure of separability between the positive and negative classes and helps to identify how much the models are capable of distinguishing between low and high scores (in our case). Accuracy evaluates how many times the models were correct overall, precision has to do with how well the models do at predicting a particular class, and finally recall shows how many times the models were able to detect a particular class. As these metrics have their caveats (e.g., accuracy is not appropriate for imbalanced datasets), we considered them all to account for various aspects of the models.

5.1. Validation phase: Single-step prediction

We evaluated the performance of the models using the four metrics and the results are illustrated in Tables 2 and 3. The tables show performance measures of the models on data from 305 learners (with 5272 solutions) using 10-fold cross-validation. Fig. 3(a) to Fig. 3(f) demonstrates predictions of the models on 20-70% of learners' actions using the AUC metric, and Fig. 4 shows their accuracy, precision, and recall on 20-50% of learners' actions. Because our proposed approach focuses on the early prediction of final scores, more attention is given to the shorter action sequences (i.e., 20-50%).

Overall, the CNN model has performed better than other models regardless of the length of action sequences. After the CNN, the LSTM model appeared to perform better compared to the RNN and MLP. Specifically, based on the AUC metric, the CNN model outperformed other models using 20-70% of the learner's actions. As shown in Table 3, according to accuracy, precision, and recall, on 20-40% and 70% of learners' actions the CNN model outperformed all other models. However, based on these metrics, on 50% and 60% of actions, the LSTM slightly performed better than others. In general, both the RNN and the baseline models showed lower performance than the other two models. Given these results, it can be concluded that the CNN model has a better prediction performance for the early prediction task, especially when it comes to the shorter length of action sequences.

Table 2
AUC measure of the models on single-step prediction using cross-validation

Length of action sequences (%)	LSTM	RNN	CNN	MLP (baseline)
20% of actions	0.762	0.745	0.805	0.698
30% of actions	0.801	0.773	0.828	0.777
40% of actions	0.766	0.799	0.869	0.777
50% of actions	0.824	0.809	0.875	0.831
60% of actions	0.831	0.831	0.858	0.804
70% of actions	0.816	0.819	0.849	0.805

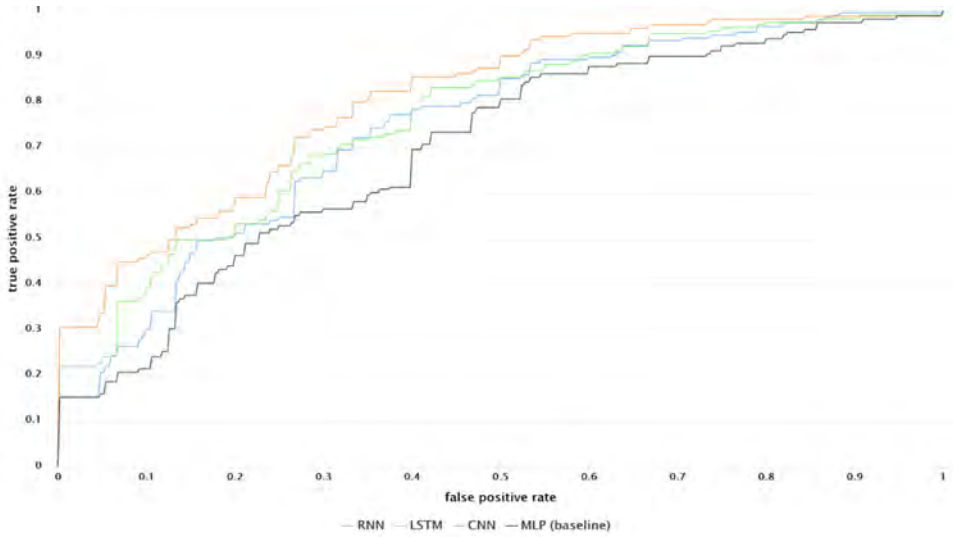


Fig. 3(a). AUC measure of the models on single-step predictions using 20% of learners' actions in cross-validation

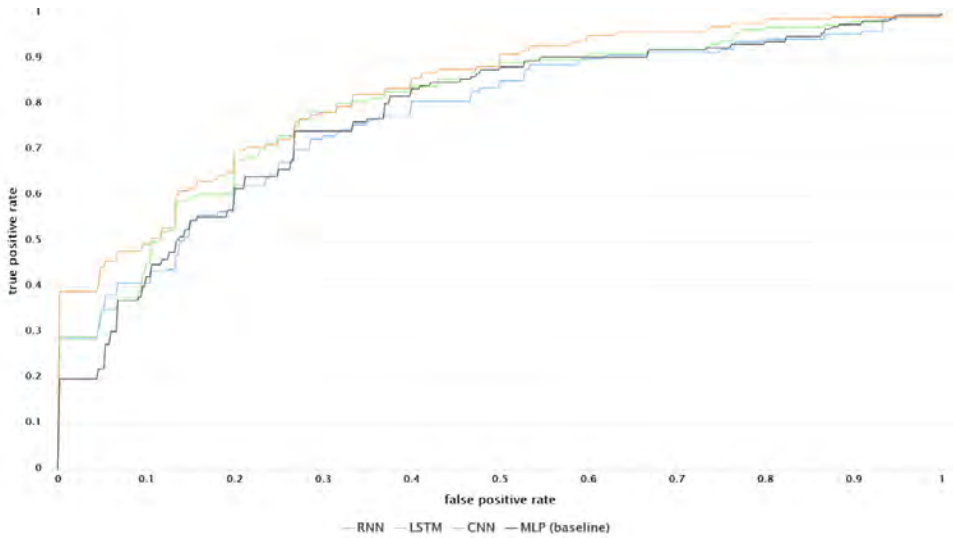


Fig. 3(b). AUC measure of the models on single-step predictions using 30% of learners' actions in cross-validation

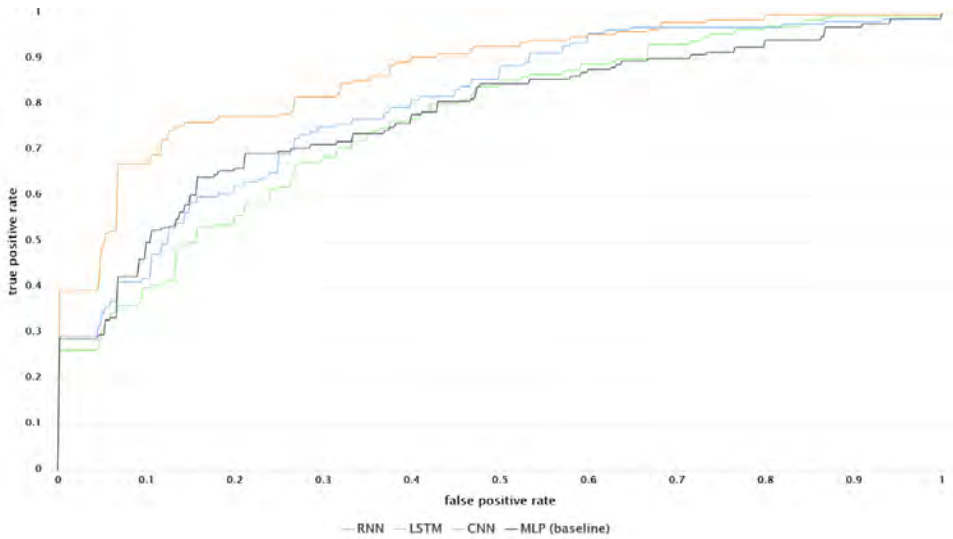


Fig. 3(c). AUC measure of the models on single-step predictions using 40% of learners' actions in cross-validation

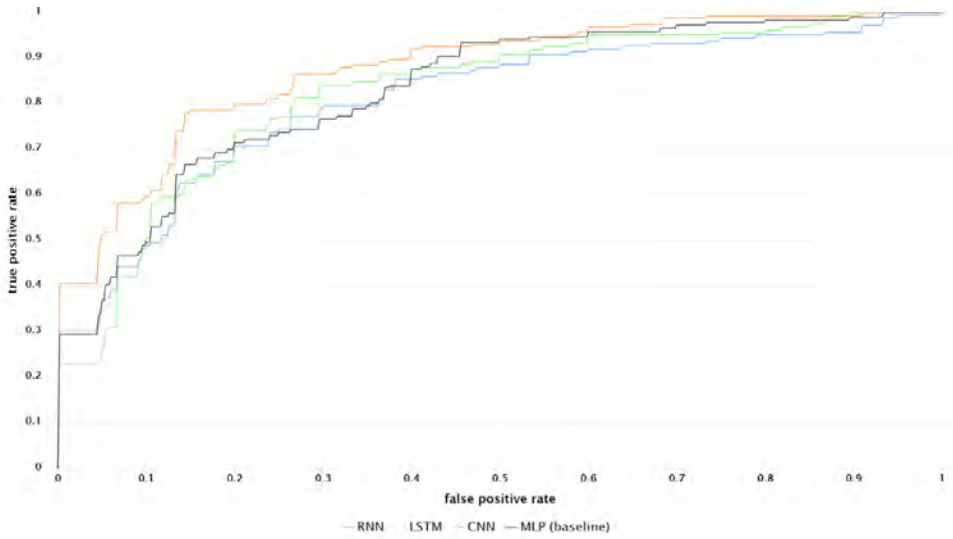


Fig. 3(d). AUC measure of the models on single-step predictions using 50% of learners' actions in cross-validation

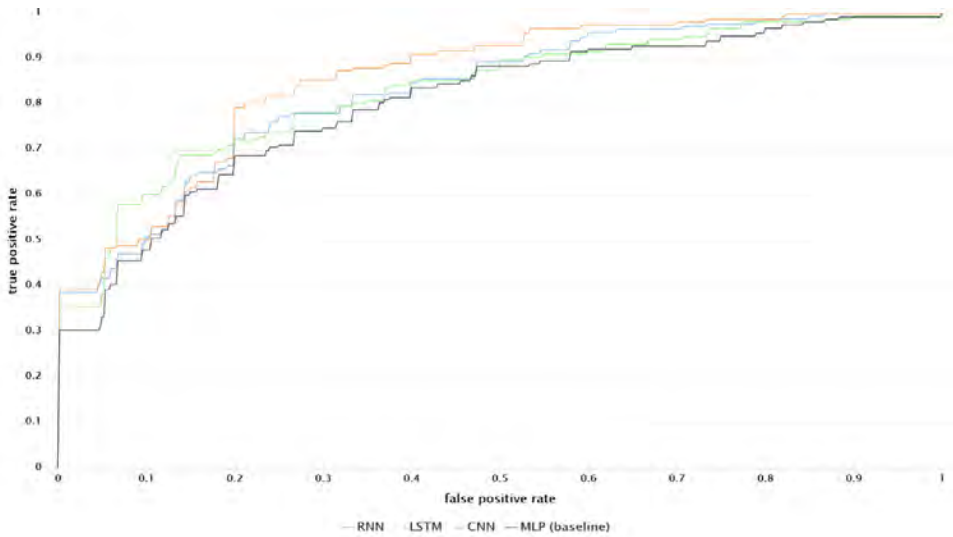


Fig. 3(e). AUC measure of the models on single-step predictions using 60% of learners' actions in cross-validation

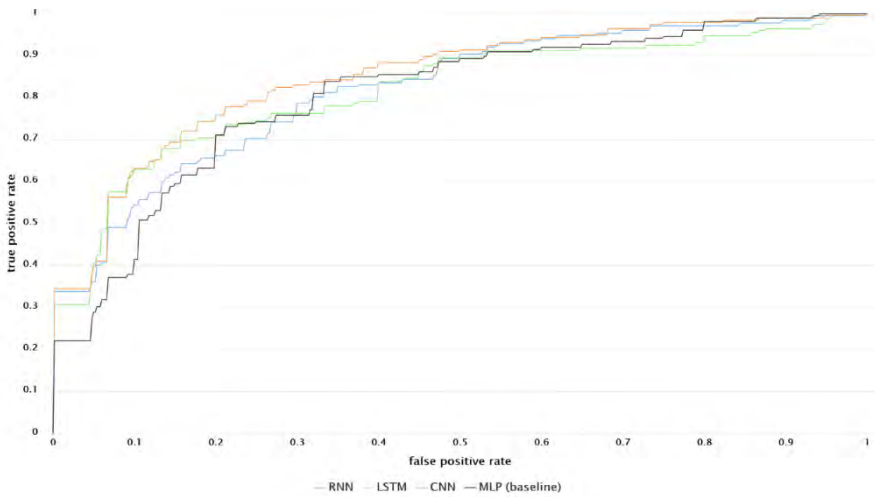


Fig. 3(f). AUC measure of the models on single-step predictions using 70% of learners' actions in cross-validation

Table 3
Model performances on single-step prediction using cross-validation

Models	Length of action sequences (%)	Accuracy (%)	Recall (%)	Precision (%)
LSTM	20	68.5	75.1	66.2
	30	72.8	78.1	70.9
	40	67.6	65.5	69.3
	50	75.2	79.6	72.3
	60	75.7	75.7	76.2
	70	74.5	77	73.2
RNN	20	66.4	68.9	66.6
	30	66.6	70.9	67.4
	40	68.4	76.3	67.4
	50	71.9	74.5	70.5
	60	72.2	76.1	71.2
	70	75.4	75.6	74.8
CNN	20	71.1	81.2	67.8
	30	73.4	81.6	70.5
	40	76	71.8	82.1
	50	72.9	77.7	74.2
	60	71.7	76.5	72.8
	70	77.5	81.9	75.3
MLP (baseline)	20	62	69.6	61.7
	30	70.5	74.8	69.9
	40	64.1	69.6	66.41
	50	72.3	73.3	71.4
	60	72.5	73.8	71.2
	70	74.3	76.8	73.4

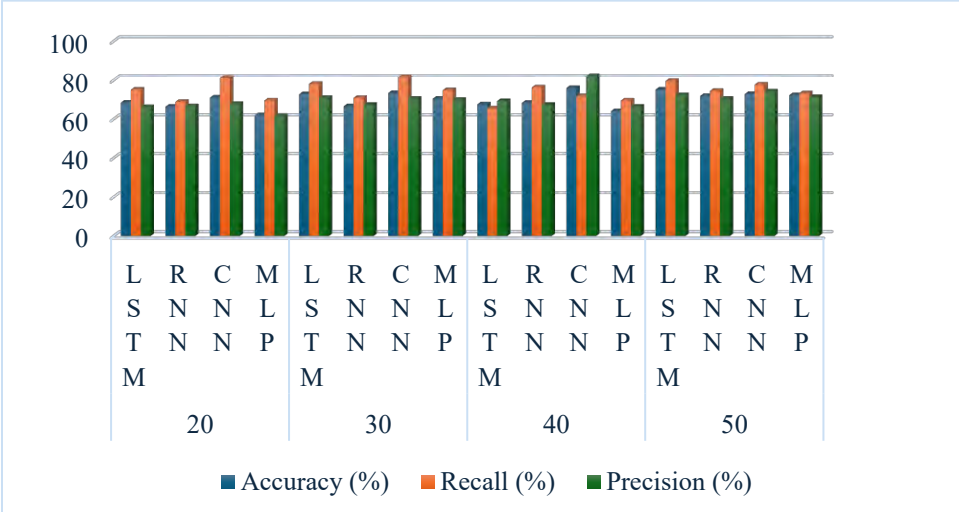


Fig. 4. Performance of the models on single-step prediction using cross-validation (accuracy, precision, and recall)

5.2. Application phase: Single-step prediction

In this section, we investigate the performance of the cross-validated models on the independent game data from 34 learners (with 586 solutions). Table 4 shows the performance of the models using the AUC, and Table 5 lists the model performance using accuracy, precision, and recall. Additionally, Fig. 5 demonstrates the models' accuracy, recall, and precision for 20-50% of action sequences.

Based on the results of Table 4, overall, the CNN model has outperformed all other models. More specifically, on 20 to 40% of action sequences, the AUC metric has selected CNN as the best. On 50%, 60%, and 70% of actions, according to the AUC metric, the MLP, LSTM, and RNN are performing slightly better, respectively. Based on the accuracy, recall, and precision, the CNN model steadily shows a better performance than the other models on 20-50% of action sequences. On 70% of action sequences, the RNN model outperformed other models based on accuracy, while the CNN model appeared to have better recall and precision. Given these results, it can be concluded that the CNN model has a better prediction performance for the early prediction task regardless of the length of action sequences.

Table 4
AUC measure of the models on single-step prediction using the independent dataset

Length of action sequences (%)	LSTM	RNN	CNN	MLP (baseline)
20% of actions	0.771	0.704	0.811	0.779
30% of actions	0.807	0.821	0.879	0.843
40% of actions	0.804	0.750	0.825	0.789
50% of actions	0.850	0.850	0.864	0.911
60% of actions	0.904	0.761	0.861	0.850
70% of actions	0.793	0.811	0.804	0.807

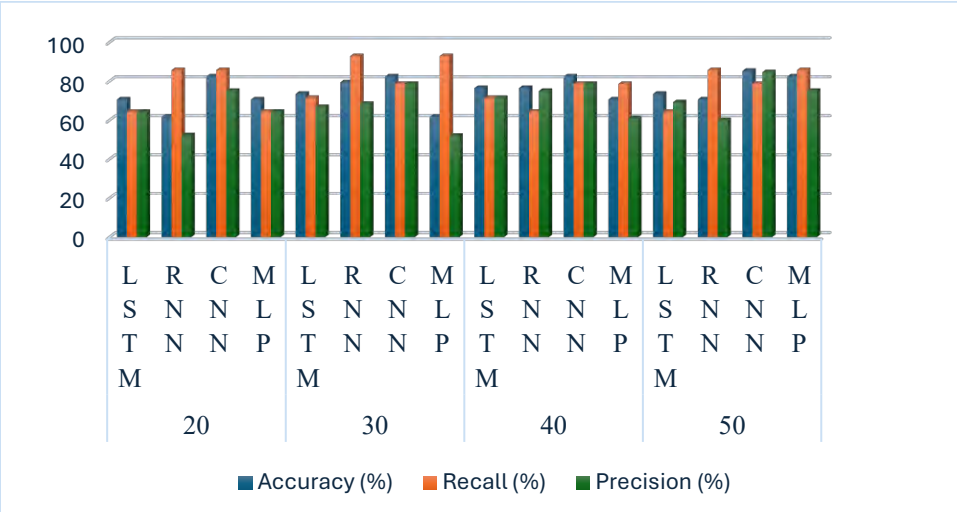


Fig. 5. Performance of the models on single-step prediction using the independent dataset (accuracy, precision, and recall)

Table 5
Model performances on single-step prediction using the independent dataset

Models	Length of action sequences (%)	Accuracy (%)	Recall (%)	Precision (%)
LSTM	20	70.6	64.3	64.3
	30	73.5	71.4	66.7
	40	76.5	71.4	71.4
	50	73.5	64.3	69.2
	60	85.3	85.7	80
	70	70.6	64.3	64.3
RNN	20	61.8	85.7	52.2
	30	79.4	92.9	68.4
	40	76.5	64.3	75
	50	70.6	85.7	60
	60	67.7	50	63.6
	70	79.4	78.6	73.3
CNN	20	82.4	85.7	75
	30	82.4	78.6	78.6
	40	82.4	78.6	78.6
	50	85.3	78.6	84.6
	60	70.6	85.7	60
	70	70.6	81.9	75.3
MLP (baseline)	20	70.6	64.3	64.3
	30	61.8	92.9	52
	40	70.6	78.6	61.1
	50	82.4	85.7	75
	60	73.5	85.7	63.2
	70	73.5	78.6	64.7

5.3. Validation and application phase: Multi-step prediction

To evaluate the performance of the deep learning models on multi-step prediction of midterm and final scores, we also used AUC, as well as average accuracy, recall, and precision metrics. Table 6 lists the AUC of the CNN and baseline model using 20-50% length of action sequences on both cross-validation (data from 5611 solutions) and independent datasets (560 solutions). The reason for comparing the baseline model with the CNN model is that in our previous experiment on single-step prediction, the CNN model was found to be the most robust model for most lengths of action sequences. Fig. 6 and Fig. 7 demonstrate predictions of the models on 20-50% of learners' actions using accuracy, recall, and precision using cross-validation and the independent dataset, respectively. Because our proposed approach early predicts both midterm and final scores at the same time, predictions of more than 50% have not been considered.

In the validation phase, regardless of the length of action sequences, the CNN consistently outperformed the MLP model on all measures. Likewise, in the application phase, the CNN model showed a better performance compared to the baseline model on all four measures. Consequently, it can be concluded that the CNN model has a better prediction performance for the early multi/step prediction task.

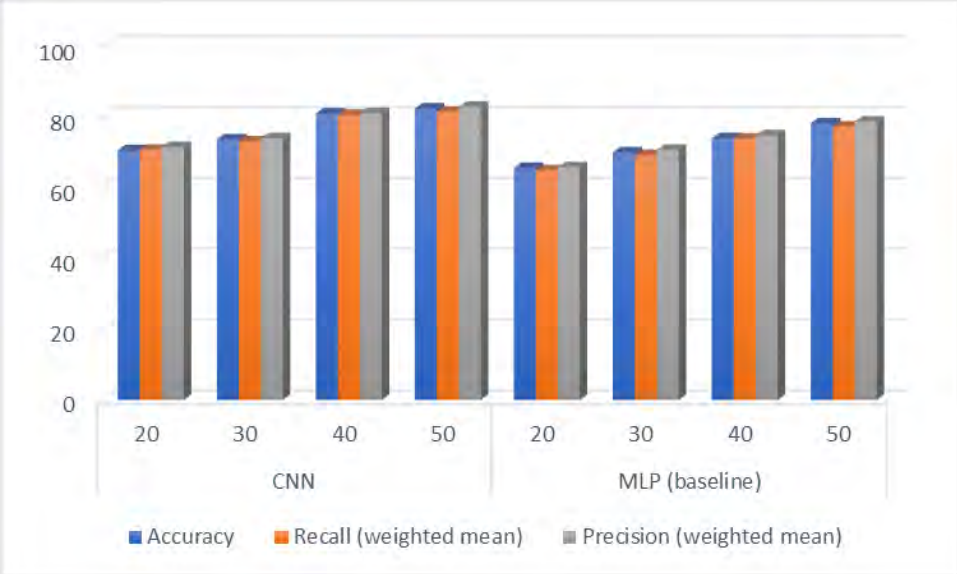


Fig. 6. Performance of the models on multi-step prediction using cross-validation (accuracy, precision, and recall)

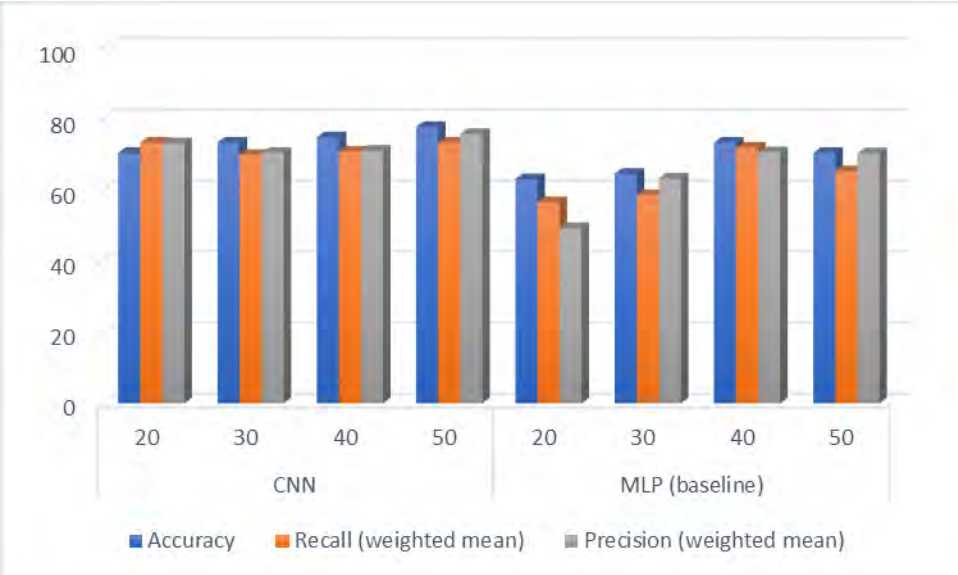


Fig. 7. Performance of the models on multi-step prediction using the independent dataset (accuracy, precision, and recall)

Table 6
AUC of the models on multi-step prediction using cross-validation and the unseen Dataset

length of action sequences (%)	CNN		MLP (baseline)	
	Cross-validation	Unseen dataset	Cross-validation	Unseen dataset
20% of actions	0.776	0.808	0.717	0.603
30% of actions	0.844	0.793	0.783	0.715
40% of actions	0.889	0.848	0.829	0.745
50% of actions	0.914	0.839	0.860	0.794

6. Discussion and conclusions

This study introduces a new method called DeepLM that uses advanced machine learning techniques to predict learners' midterm and final scores in computer-based educational games. The proposed approach considers learners' actions during gameplay and their estimated knowledge level to predict future performance. The results of this study suggest that the proposed approach can accurately predict both midterm and final scores with an AUC of up to 0.91 and 0.84, respectively, using cross-validation and independent datasets. The CNN model was also found to be effective, achieving an accuracy of 82% and 78% using cross-validation and independent datasets, respectively. Overall, the CNN model consistently outperformed other models, especially the RNN and MLP (baseline). More specifically, regarding the performance of the models on the single-step prediction of the final score in the validation phase, results showed that the CNN could robustly distinguish between the high and low scores, outperforming other models with AUC and accuracy of higher than 0.80 and 71%, respectively, in all length of action sequences. While not as high as the CNN, the LSTM model achieved performance better than both the RNN and the baseline (MLP). The lowest performance belongs to the MLP (baseline model, with accuracy as low as 62% for the 20% action sequences. Concerning the performance of different models on the single-step prediction of the final score in the application phase, results showed that the CNN could successfully predict learners' final score, especially in the shorter length of action sequences. Particularly, in sequence lengths of 20-50%, the CNN model showed better performance, while in 60% and 70% action lengths, the LSTM and RNN appeared to be performing better. Aside from 50% of action lengths, the baseline model seemed to be the lowest performing.

When it comes to early prediction of both midterm and final scores simultaneously, in both validation and application phases, the CNN model could achieve AUC ranging from 0.77 to 0.91, outperforming the baseline model. The highest and lowest AUC was achieved using 50% and 20% of action sequences, respectively. Additionally, the CNN model achieved accuracy as high as 82% on 50% of action sequences in the validation phase, as well as accuracy as high as 78% on 50% of action sequences in the application phase. Interestingly, unlike other models, the performance of the CNN model did not decrease much either by increment in the length of action sequences or by being validated against independent datasets. Therefore, the CNN model could robustly early predict the final and midterm scores at the same time using both cross-validation and unseen datasets.

Overall, a closer look at the behaviour of the models reveals that the LSTM model sometimes shows performance close to the CNN, especially in longer lengths of action

sequences. The unsurprising reason is that for a long time, the LSTM holds single observations in its memory. This may not be helpful in cases when there are smaller or shorter models that do not require remembering long-term sequential dependencies. In our findings, it was also observed that the RNN seems to be more successful using shorter lengths of action sequences because it often looks at the recent inputs (short-term modelling). Thus, the LSTM showed to have benefited from the increment in length of action sequences, whereas the RNN model's performance may fall by increment in action sequences. Finally, as the CNN model can identify local patterns better, it appears to be the most robust and best performing in comparison to other models. In other words, the CNN model seems to learn patterns within the time window better without having assumptions about the history being complete. Lastly, the MLP baseline model appeared not to be as successful as other models in handling the sequential data, and to have lower performance on shorter sequence lengths which is imperative for early predictive tasks.

Surprisingly, although CNN is not specially designed for sequential tasks and non-image data (in general), it could outperform specialized deep learning models like RNN and LSTM. The reason is that CNN considers the assumption that similar local patterns are related everywhere and ignores relationships between each sequence step's hidden vectors. Another reason that makes the CNN model stand out is its computational lightness due to fewer sequential calculations. The results of our experiments are aligned with the findings of Hooshyar et al. (2022) and Nabi et al. (2021) as they also reported the superiority of the CNN over sequential models like RNN and LSTM when predicting using small feature sizes.

From an educational perspective, the incorporation of our proposed DeepLM approach holds the potential for integrating stealth and continuous assessment of learners' knowledge and skills into various types of educational games. By employing DeepLM, we can effectively harness in-game data to enable accurate early predictions of learners' performance. For instance, when developing solution number five (out of 20 in the Autothinking game), the DeepLM approach accurately early predicts learners' performance at the end of solution number 12 and 20.

The utilization of early predictions afforded by DeepLM enables the provision of optimal learning items or task sequences tailored to individual learners. This personalized approach allows educators and game designers to offer targeted feedback, hints, and interventions that cater to learners' specific needs. Furthermore, the adaptive learning task sequences generated by DeepLM facilitate the creation of NPCs that can dynamically adapt to learners' abilities and requirements. Ultimately, these enhancements contribute to improving educational game retention by fostering a highly engaging and tailored learning experience.

It is noteworthy that our study confirms the findings of Lee-Cultura et al. (2020) that in-game data can be utilized for early prediction of learners' performance. However, unlike Lee-Cultura et al. (2020) work, our proposed approach employs a state-of-the-art deep neural network to model latent information from in-game data. Additionally, our approach demonstrates efficacy in both single- and multi-predictions of learners' midterm and final scores simultaneously. This versatility offers educators and game designers a comprehensive understanding of learners' progress and enables them to make informed decisions regarding instructional strategies and interventions.

6.1. Limitations and future works

There are a number of limitations in the proposed approach that need to be addressed in future works. Although our proposed approach could achieve high accuracy while maintaining a relatively balanced trade-off between recall and precision, yet mis-classifies some computational thinking solutions. Considering the high-risk nature of education, future research can explore the implementation of hybrid models and employ advanced data augmentation techniques to enhance the accuracy of predictions. Furthermore, it is advantageous to apply the suggested method to datasets from additional educational games in order to more thoroughly explore its applicability across various contexts. Finally, as the effectiveness of the proposed approach on learners in the real-world is unknown, it is desirable to integrate the proposed approach into the game and evaluate its effectiveness in real-world classrooms, providing personalized learning that is interpretable.

Author Statement

The authors declare that there is no conflict of interest.

Acknowledgements

This work was supported by the Tallinn University project entitled "Fostering the research strand in Artificial Intelligence in Education at TLU" with number TF/1422, the COPCOT (ANR-22-CE38-0003) project funded by the Agence Nationale de Recherche, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1C1C2004868).

ORCID

Danial Hooshyar https://orcid.org/0000-0002-9143-6648

Nour El Mawas https://orcid.org/0000-0002-0214-9840

Yeongwook Yang https://orcid.org/0000-0003-3219-7250

References

Abdul Jabbar, A. I., & Felicia, P. (2015). Gameplay engagement and learning in game-based learning: A systematic review. *Review of Educational Research, 85*(4), 740–779. https://doi.org/10.3102/0034654315577210

Akram, B., Min, W., Wiebe, E., Mott, B., Boyer, K. E., & Lester, J. (2018, July). Improving stealth assessment in game-based learning with LSTM-based analytics. In *Proceedings of the 2018 International Conference on Educational Data Mining* (pp. 208–218). National Science Foundation. Retrieved from https://par.nsf.gov/servlets/purl/10100664

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4*(2), 167–207. https://doi.org/10.1207/s15327809jls0402_2

Bakan, U., Han, T., & Bakan, U. (2022). Learner perceptions and effectiveness of using a

massively multiplayer online role-playing game to improve EFL communicative competence. *Knowledge Management & E-Learning, 14*(3), 286–303. https://doi.org/10.34105/j.kmel.2022.14.016

Blakely, G., Skirton, H., Cooper, S., Allum, P., & Nelmes, P. (2009). Educational gaming in the health sciences: Systematic review. *Journal of Advanced Nursing, 65*(2), 259–269. https://doi.org/10.1111/j.1365-2648.2008.04843.x

Brezovszky, B., McMullen, J., Veermans, K., Hannula-Sormunen, M. M., Rodríguez-Aflecht, G., Pongsakdi, N., Laakkonen, E., & Lehtinen, E. (2019). Effects of a mathematics game-based learning environment on primary school students' adaptive number knowledge. *Computers & Education, 128*, 63–74. https://doi.org/10.1016/j.compedu.2018.09.011

Chen, C. H., & Law, V. (2016). Scaffolding individual and collaborative game-based learning in learning performance and intrinsic motivation. *Computers in Human Behavior, 55*(Part B), 1201–1212. https://doi.org/10.1016/j.chb.2015.03.010

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253–278. https://doi.org/10.1007/BF01099821

El Mawas, N., Hooshyar, D., & Yang, Y. (2020b, May). Investigating the learning impact of autothinking educational game on adults: A case study of France. In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020) (Vol. 2)* (pp. 188–196). SCITEPRESS Digital Library. Retrieved from https://nour-elmawas.com/publis/CSEDUElMawasTartu2020.pdf

El Mawas, N., Tal, I., Moldovan, A. N., Bogusevschi, D., Andrews, J., Muntean, G. M., & Muntean, C. H. (2020a). Investigating the impact of an adventure-based 3D solar system game on primary school learning process. *Knowledge Management & E-Learning, 12*(2), 165–190. https://doi.org/10.34105/j.kmel.2020.12.009

Emerson, A., Rodríguez, F. J., Mott, B., Smith, A., Min, W., Boyer, K. E., Smith, C., Wiebe, E., & Lester, J. (2019, July). Predicting early and often: Predictive student modeling for block-based programming environments. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)* (pp. 39–48). International Educational Data Mining Society. Retrieved from https://files.eric.ed.gov/fulltext/ED599223.pdf

Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., & Lester, J. (2021). Predictive student modeling in game-based learning environments with word embedding representations of reflection. *International Journal of Artificial Intelligence in Education, 31*(1), 1–23. https://doi.org/10.1007/s40593-020-00220-4

Ha, E. Y., Rowe, J. P., Mott, B. W., & Lester, J. C. (2012, July). Goal recognition with Markov logic networks for player-adaptive games. *Proceedings of the AAAI Conference on Artificial Intelligence, 26*(1), 2113–2119. https://doi.org/10.1609/aaai.v26i1.8439

Henderson, N., Kumaran, V., Min, W., Mott, B., Wu, Z., Boulden, D., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., & Lester, J. (2020, July). 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)* (pp. 92–103). International Educational Data Mining Society. Retrieved from https://files.eric.ed.gov/fulltext/ED607823.pdf

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. *Complexity*. https://doi.org/10.1155/2019/1306039

Hooshyar, D. (2022). Effects of technology-enhanced learning approaches on learners with

different prior learning attitudes and knowledge in computational thinking. *Computer Applications in Engineering Education, 30*(1), 64–76. https://doi.org/10.1002/cae.22442

Hooshyar, D., Huang, Y. M., & Yang, Y. (2022). GameDKT: Deep knowledge tracing in educational games. *Expert Systems with Applications, 196*: 116670. https://doi.org/10.1016/j.eswa.2022.116670

Hooshyar, D., Lim, H., Pedaste, M., Yang, K., Fathi, M., & Yang, Y. (2019b, November). AutoThinking: An adaptive computational thinking game. In *Proceedings of the Second International Conference on Innovative Technologies and Learning (ICITL 2019)* (pp. 381–391). Springer. https://doi.org/10.1007/978-3-030-35343-8_41

Hooshyar, D., Malva, L., Yang, Y., Pedaste, M., Wang, M., & Lim, H. (2021b). An adaptive educational computer game: Effects on students' knowledge and learning attitude in computational thinking. *Computers in Human Behavior, 114*: 106575. https://doi.org/10.1016/j.chb.2020.106575

Hooshyar, D., Pedaste, M., Yang, Y., Malva, L., Hwang, G. J., Wang, M., Lim, H., & Delev, D. (2021a). From gaming to computational thinking: An adaptive educational computer game-based learning approach. *Journal of Educational Computing Research, 59*(3), 383–409. https://doi.org/10.1177/0735633120965919

Hooshyar, D., Yousefi, M., & Lim, H. (2019a). A systematic review of data-driven approaches in player modeling of educational games. *Artificial Intelligence Review, 52*(3), 1997–2017. https://doi.org/10.1007/s10462-017-9609-8

Jimenez, F., Paoletti, A., Sanchez, G., & Sciavicco, G. (2019). Predicting the risk of academic dropout with temporal multi-objective optimization. *IEEE Transactions on Learning Technologies, 12*(2), 225–236. https://doi.org/10.1109/TLT.2019.2911070

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying evidence-centered design for the development of game-based assessments in physics playground. *International Journal of Testing, 16*(2), 142–163. https://doi.org/10.1080/15305058.2015.1108322

Lee-Cultura, S., Sharma, K., Papavlasopoulou, S., & Giannakos, M. (2020, August). Motion-based educational games: Using multi-modal data to predict player's performance. In *Proceedings of the 2020 IEEE Conference on Games (CoG)* (pp. 17–24). IEEE. https://doi.org/10.1109/CoG47356.2020.9231892

Liu, T. (2022). Knowledge tracing: A bibliometric analysis. *Computers and Education: Artificial Intelligence, 3*: 100090. https://doi.org/10.1016/j.caeai.2022.100090

Liu, Z., Moon, J., Kim, B., & Dai, C. P. (2020). Integrating adaptivity in educational games: A combined bibliometric analysis and meta-analysis review. *Educational Technology Research and Development, 68*(4), 1931–1959. https://doi.org/10.1007/s11423-020-09791-4

Malva, L., Hooshyar, D., Yang, Y., & Pedaste, M. (2020, July). Engaging Estonian primary school children in computational thinking through adaptive educational games: A qualitative study. In *Proceedings of the 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)* (pp. 188–190). IEEE. https://doi.org/10.1109/ICALT49669.2020.00061

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E., & Lester, J. C. (2019). DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies, 13*(2), 312–325. https://doi.org/10.1109/TLT.2019.2922356

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Morshed Fahid, F., Tian, X., Emerson, A., B. Wiggins, J., Bounajim, D., Smith, A., Wiebe, E., Mott, B., Elizabeth Boyer, K., & Lester, J. (2021). Progression trajectory-based student modeling for novice block-based programming. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21)* (pp. 189–200). Association for Computing Machinery. https://doi.org/10.1145/3450613.3456833

Nabi, K. N., Tahmid, M. T., Rafi, A., Kader, M. E., & Haider, M. A. (2021). Forecasting COVID-19 cases: A comparative analysis between recurrent and convolutional neural networks. *Results in Physics, 24*: 104137. https://doi.org/10.1016/j.rinp.2021.104137

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences, 11*(1): 237. https://doi.org/10.3390/app11010237

Olive, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2019). A quest for a one-size-fits-all neural network: Early prediction of students at risk in online courses. *IEEE Transactions on Learning Technologies, 12*(2), 171–183. https://doi.org/10.1109/TLT.2019.2911068

Ongoro, C. A., & Mwangoka, J. W. (2019). Effects of digital games on enhancing language learning in Tanzanian preschools. *Knowledge Management & E-Learning, 11*(3), 325––344. https://doi.org/10.34105/j.kmel.2019.11.017

Parsons, J., & Taylor, L. (2011). Improving student engagement. *Current Issues in Education, 14*(1). Retrieved from https://cie.asu.edu/ojs/index.php/cieatasu/article/view/745

Pesare, E., Roselli, T., Corriero, N., & Rossano, V. (2016). Game-based learning and gamification to promote engagement and motivation in medical learning contexts. *Smart Learning Environments, 3*: 5. https://doi.org/10.1186/s40561-016-0028-0

San Pedro, M. O. Z., d Baker, R. S., Gowda, S. M., & Heffernan, N. T. (2013, July). Towards an understanding of affect and knowledge from student interaction with an intelligent tutoring system. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED 2013)* (pp. 41–50). Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-39112-5_5

Shute, V. J., & Kim, Y. J. (2012). Does playing World of Goo facilitate learning? In D. Y. Dai (Ed.), *Design Research on Learning and Thinking in Educational Settings: Enhancing Intellectual Growth and Functioning* (pp. 252–276). Routledge.

Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In J. Spector, M. Merrill, J. Elen, & M. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 311–321). Springer.

Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing and supporting competencies within game environments. *Technology, Instruction, Cognition and Learning, 8*(2), 137–161.

Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games.* The MIT press.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

Vlachopoulos, D., & Makri, A. (2017). The effect of games and simulations on higher education: A systematic literature review. *International Journal of Educational Technology in Higher Education, 14*: 22. https://doi.org/10.1186/s41239-017-0062-1

Wang, L., Sy, A., Liu, L., & Piech, C. (2017, April). Deep knowledge tracing on programming exercises. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale (L@S '17)* (pp. 201–204). Association for Computing Machinery.

Retrieved from https://dl.acm.org/doi/10.1145/3051457.3053985

Xie, J., Wang, M., & Hooshyar, D. (2021). Student, parent, and teacher perceptions towards digital educational games: How they differ and influence each other. *Knowledge Management & E-Learning, 13*(2), 142–160. https://doi.org/10.34105/j.kmel.2021.13.008

Yannakakis, G. N., & Togelius, J. (2018). Modeling players. In G. N. Yannakakis & J. Togelius (Eds.), *Artificial Intelligence and Games* (pp. 203–255). Springer. https://doi.org/10.1007/978-3-319-63519-4_5