Enhancing Student Competency Models for Game-Based Learning with a Hybrid Stealth Assessment Framework

Nathan Henderson[1], Vikram Kumaran[1], Wookhee Min[1], Bradford Mott[1], Ziwei Wu[1],
Danielle Boulden[1], Trudi Lord[2], Frieda Reichsman[2], Chad Dorsey[2], Eric Wiebe[1], James Lester[1]

[1]North Carolina State University
Raleigh, NC, 27695
{nlhender, vkumara, wmin, bwmott, zwu17, dmboulde, wiebe, lester}@ncsu.edu

[2]Concord Consortium
Concord, MA, 01742
{tlord, freichsman, cdorsey}@concord.org

ABSTRACT

In recent years, game-based learning has shown significant promise for creating engaging and effective learning experiences. Developing models that can predict whether students will struggle with mastering certain concepts could guide adaptive support to assist students with mastering those concepts. Game-based learning environments offer significant potential for unobtrusively assessing student learning without interfering with gameplay through stealth assessment. Prior work on stealth assessment has focused on a single machine learning technique such as dynamic Bayesian networks or long short-term memory networks; however, a single modeling technique often does not guarantee the best predictive performance for all concepts of interest. In this paper, we present a hybrid data-driven approach to stealth assessment for predicting students' mastery of concepts through interactions with a game-based learning environment for introductory genetics. Stealth assessment models utilize students' observed gameplay behaviors using challenge- and session-based features to predict students' learning outcomes on identified concepts. We present single-task and multi-task models for predicting students' mastery of concepts and the results suggest that the hybrid stealth assessment framework outperforms individual models and holds significant potential for predicting student competencies.

Keywords

Stealth Assessment, Predictive Student Modeling, Game-based Learning, Multi-Task Learning

1. INTRODUCTION

Recent years have seen growing interest in game-based learning environments because of their potential for creating engaging and effective learning experiences [7, 43]. Researchers have investigated game-based learning environments in a wide array of domains, including mathematics [13, 40], computational thinking [4, 16], and science [2, 10, 30].

While a common gameplay design adopted by many game-based learning environments is providing students with a fixed sequence of levels with increasingly difficult challenges per concept, game-based learning environments could provide individualized sequences of challenges and just-in-time support, so that students can focus on gameplay at the edge of their knowledge and skills and remain engaged throughout the learning experience [20, 44]. To achieve this goal, game-based learning environments should be equipped with the ability to detect when students are struggling or have gaps in their knowledge and take appropriate action to tailor their learning experience [33]. Presenting in-game challenges adaptively tailored to individual students' knowledge can play a crucial role in supporting mastery learning and promoting engagement while effectively addressing problems with a one-size-fits-all approach.

With recent advances in machine learning, data-driven approaches using students' in-game behaviors have enabled the automatic assessment of students' evolving competence [1, 21] and the modeling of mind wandering [19], wheel spinning [25], and quitting behaviors [12], all of which are associated with negative learning outcomes. A robust modeling of student behaviors can guide students from undertaking a challenge that is beyond their capabilities as well as facilitate their engagement through individualized learning activities tailored to their competencies for knowledge and game-playing skills.

There is now a sizable literature on stealth assessment in game-based learning [34]. Stealth assessment robustly measures student learning without disrupting engagement by embedding unobtrusive assessments within game mechanics, offering real-time non-disruptive assessment [35]. Building on evidence-centered design (ECD) [24], which provides a systematic approach to developing knowledge assessments, stealth assessment examines student interaction data (i.e., evidence model) with in-game challenges (i.e., task model) to provide real-time behind-the-scenes measurement of students' learning processes and outcomes (i.e., competency model) [22, 36]. Specifically, students' learning is inferred by analyzing low-level sequences of observed problem-solving behaviors that manifest competencies for knowledge and skills without conducting explicit formative assessments. Inferences made by stealth assessment models can inform effective

Nathan Henderson, Vikram Kumara, Wookhee Min, Bradford Mott, Ziwei Wu, Danielle Boulden, Trudi Lord, Frieda Reichsman, Chad Dorsey, Eric Wiebe and James Lester "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),* Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 92 - 103

scaffolding strategies (e.g., adaptive challenge selection, tailored problem-solving support) for individual learners in a timely and contextually appropriate manner [29, 36]. It can also guide teachers to potential pedagogical adaptations or support integration with additional curricular activities, which are core components of distributed and integrated scaffolding [26, 28, 39].

In this work, we investigate stealth assessment with *Geniventure*, a game-based learning environment for introductory high school genetics learning. We present modeling approaches including single-task and multi-task random forest and recurrent neural network-based regression models for predicting students' competencies, whose labels were derived from students' post-test scores on genetics-focused concepts. In contrast to much previous work on stealth assessment that used a single machine learning technique, we present a hybrid stealth assessment framework that effectively leverages predictive capacities of all the explored modeling approaches. We compare the models' fitness to the data to gain insight into which combinations of models perform optimally across all the concepts, as well as which models are effective for individual concepts. The hybrid stealth assessment framework outperforms individual computational techniques with respect to predictive performance for student concept-level competencies.

2. RELATED WORK

Intelligent game-based learning environments simultaneously leverage capabilities of digital games to motivate students' learning through engaging narratives, virtual environments and intelligent tutoring systems (ITSs) to foster students' learning through adaptive scaffolding and context-sensitive feedback [15]. These environments facilitate learning through individualized challenges, narratives, feedback, and problem-solving support [30, 35, 42]. Students' fine-grained, sequential game trace data has been used in a wide range of student modeling tasks such as inferring the level of competency [22, 35], predicting affective states [3, 31], and recognizing students' learning goals [23]. In comparison to single-task learning investigated in much of previous student modeling work, recent years have seen a growing interest in the use of multi-task learning, a regularization method that exploits commonalities and differences across related tasks for improved generalizability. Multi-task learning has been examined for various student modeling tasks such as predicting student competencies in programming in a massive open online course (MOOC) [27] and modeling student performance in a game-based learning environment for middle-grade microbiology education [9], which demonstrated improved predictive performance relative to the single-task modeling approach. Similarly, Chaudhry et al. used multi-task modeling with both hint usage and knowledge tracing to induce models of students using online tutoring systems [5].

Stealth assessment is methodologically grounded in evidence-centered design (ECD), which was proposed to construct educational assessments in terms of evidentiary arguments [24]. ECD features task, evidence, and competency models to conduct probabilistic reasoning about knowledge, skills, and abilities of students utilizing evidence captured from interactions with learning tasks. Stealth assessment conducts real-time processing of data derived from these three ECD models that informs intelligent, adaptive game-based learning environments through devising robust evidence and competency models as well as creating task models that effectively develop the competencies [20]. While human expert-designed Bayesian networks have been examined as the core computational method for both competency and evidence models for stealth assessment [37], another body of work has investigated an assessment pipeline that does not require costly domain knowledge engineering. Falakmasir et al. investigate the use of hidden Markov models (HMMs) to model student proficiency within educational games [8]. The log-likelihoods are approximated by the HMMs using sequential gameplay data, with the difference between the likelihoods serving as the independent variable for post-test prediction models. The authors of this work use linear regression to predict the student's post-test scores. There has also been growing interest in deep neural network architectures due to their capability to learn salient features from low-level, sequential data captured from interactions with task models [1, 20]. Long short-term memory network-based stealth assessment models have demonstrated significant promise by outperforming competitive baselines with respect to predictive performance of inferring students' competencies, while effectively eliminating the need to manually craft evidence rules and evidence models. In contrast to much of previous research, our work presents a hybrid stealth assessment framework that utilizes a suite of competency models to optimally harness distinguished predictive capacity yielded by a range of single-task and multi-task stealth assessment models.

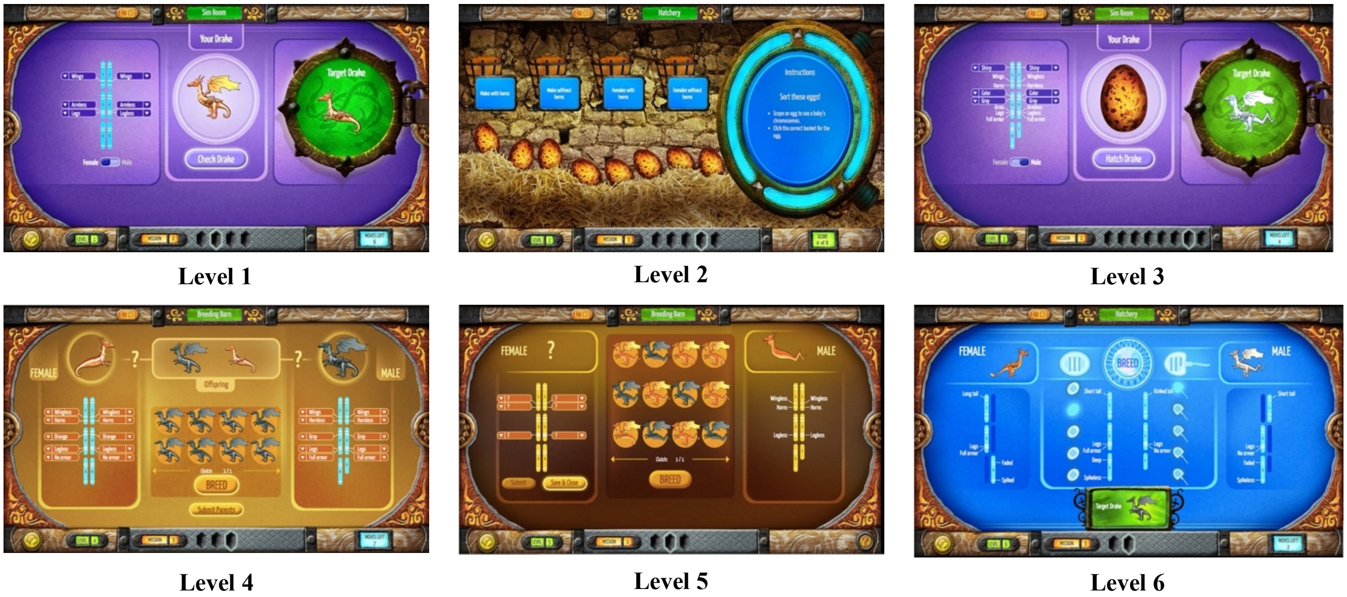
3. DATASET

3.1 *Geniventure* Learning Environment

To evaluate the performance of our hybrid stealth assessment framework, we use gameplay interaction log data collected from students engaged with a game-based learning environment for introductory genetics for middle school and high school students (students ages 11-18 years), *Geniventure*. The design of the game is guided by core genetics-based concepts that align with the Next Generation Science Standards [38]. *Geniventure* engages students in exploring heredity, dominant and recessive traits, and the protein-to-trait relationship by breeding and studying drakes, a model species for dragons [18].

The game consists of 60 increasingly difficult puzzle-like challenges across 6 levels (Figure 1). Each of the challenges is part of a "mission", with each level containing multiple missions. The genetics concepts that the game addresses are presented through a variety of challenge types. While the game was designed to be played through in a linear fashion, students have the freedom to attempt challenges at any level and are allowed to quit a challenge at any time.

In the first half of the game, students are asked to change the drake's genotype to match a target phenotype (Figure 1, Level 1). To successfully complete these problem-solving challenges, students must understand several genetic concepts and be able to infer the phenotype of their drake from its genotype. Once students feel they have the correct genotype, they click the "Check" or "Hatch" button to submit their answer. If the drake they create matches the target drake, the challenge is successfully completed. Otherwise, the game provides the student with three progressively more directed levels of hints, as well as a visual cue, and allows them to continue to make further changes to the alleles until they quit or successfully complete the challenge. This model of counting moves and giving feedback in the form of hints is carried through the subsequent levels of the game, even though the challenge types vary. Other challenges instruct the user to match a phenotype to a given genotype, following a reversed procedure from Level 1 (Figure 1, Level 2), and also introduce scale color and other additional complexities to the challenges (Figure 1, Level 3).



Level 1 **Level 2** **Level 3**

Level 4 **Level 5** **Level 6**

Figure 1. Example challenges in *Geniventure* for the six gameplay levels.

The latter half of the game introduces more difficult concepts such as breeding and inheritance. Through several scaffolded challenges, students breed parent drakes with the goal of matching target offspring (Figure 1, Level 4). The tasks grow increasingly complex as students progress through this level, eventually culminating in a challenge requiring students to breed two parents to produce offspring that match a given drake. Students are also introduced to test cross, a genetic method for determining the genotype of one organism by crossing it with a fully recessive organism (Figure 1, Level 5). Finally, the last level introduces traits with more complex inheritance patterns, such as X-linked and polyallelic traits (Figure 1, Level 6). This level contains challenges illustrating concepts from all of the preceding levels such as allele target match, egg drop, meiosis, breeding, and test cross.

As previously mentioned, students can validate their work at any time and are provided with system-generated hints based on their perceived understanding of the genetics concepts if necessary. Hint usage, as well as time spent on challenges, and the students' success rate during their respective gameplay sessions, serves as the foundation for the features used to train the competency models.

3.2 Data Collection

The dataset was collected from 462 students from seven high schools and one middle school located in the Middle to Northern Atlantic coast of the United States. This data was collected during a teacher-led classroom implementation of *Geniventure* where students played the game during class over the course of several days. During gameplay, students' gameplay trajectory and their detailed in-game actions were recorded as trace data logs. Before playing the game, students took a pre-test consisting of 28 questions related to the genetic concepts covered in the game. Once gameplay concluded, students took a post-test which was identical to the pre-test (Figure 2). This assessment was aligned to the ECD competency model of the game and previously validated through two rounds of expert review and cognitive interviews with students. In administration, it demonstrated an internal consistency reliability of alpha = 0.873. Both the pre-test and post-test were online surveys accessible through the same online portal as the game. 38 students were removed due to the partial or missing pre/post test data. 108 students were removed due to missing trace data, resulting in a dataset containing trace data from 316 students. Results from a paired t-test on students' knowledge pre-test (M = 14.41, SD = 5.826) and post-test (M = 19.33, SD = 6.131) revealed a significant improvement from pre-test to post-test (t(315) = 14.663, $p < 0.01$, Cohen's d = 0.823). A majority of the students attempted between 50 and 150 challenges. The fewest number of challenges attempted by a student was 5, which serves as the basis for the sequence length of the subsampling window used to generate the sequential data for the competency models. The most challenges attempted by a student during the duration of the study is 248. To further illustrate the distribution of the number of challenges attempted per student, a histogram of the students' gameplay trajectories is shown in Figure 3.

4. ECD FOR STEALTH ASSESSMENT

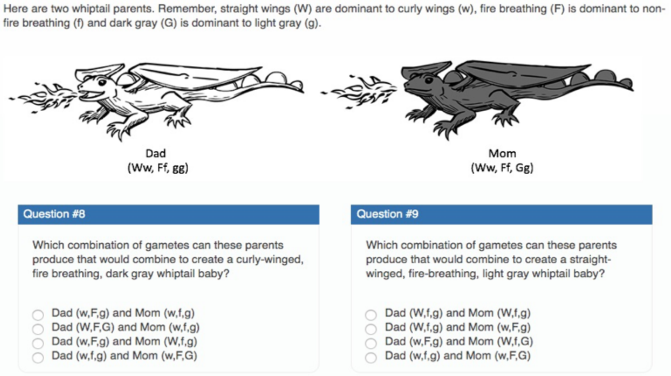
Evidence-centered design (ECD) is a systematic approach for designing and developing reliable knowledge assessments in terms of evidentiary arguments [24]. When utilized to identify and analyze user behavior in online learning environments, it serves as the basis of stealth assessment in game-based learning environments [34]. While historically ECD has been utilized in the development of summative assessments, recent years have seen its application in the design of formative stealth assessment models for game-based learning environments [20, 34]. Assessment results inferred by stealth assessment models can be utilized to support student learning through adaptive scaffolding within the learning environment and also inform teachers about student learning trajectories through a teacher dashboard. As noted above, stealth assessment is grounded in three core ECD models. These three models were applied to the current study using *Geniventure* as follows:

- ***Task Model:*** This model defines the activities, or tasks, that students undertake as part of their learning. In the *Geniventure* learning environment, the task model consisted of 60 challenges across six game levels that students undertake.

These tasks focus on genetics concepts such as heredity, dominance/recessive, and the protein-to-trait relationship.

- ***Evidence Model:*** The evidence model takes as input low-level action sequences students produce while interacting with the game-based learning environment. Game-based learner behaviors are linked to targeted concepts to generate machine-interpretable evidence that can be directly utilized with the modeling techniques presented here. That is, a probabilistic model is constructed from analysis of a series of actions related to mastery (or not) of a particular concept. The evidence model informs the competency model in order to update its belief of students' competencies as they interact with the tasks.
- ***Competency Model:*** Mastery of 16 concepts (Table 1) are dynamically estimated by the competency model with respect to students' genetics knowledge. The concepts were derived from expert review of classroom learning goals and state science standards. The ground truth for their summative competencies are acquired from students' post-test scores on an explicit content knowledge assessment. The competency model is aligned to the summative post-test through the same set of ECD-derived concepts in Table 1.

In training the stealth assessment models, we extract competency scores based on correctness of students' individual responses to items on a post-test knowledge assessment (Figure 2). Competencies for a single concept in our competency model can be evaluated in as few as one or as many as six items on the post-test survey since an assessment item can map to either one or two concepts. Item 28 is an open-ended question that can be answered in many unique ways, so we omit it from our competency score calculations. The mappings from each concept to individual survey questions can be found in Table 1.



Here are two whiptail parents. Remember, straight wings (W) are dominant to curly wings (w), fire breathing (F) is dominant to non-fire breathing (f) and dark gray (G) is dominant to light gray (g).

Dad
(Ww, Ff, gg)

Mom
(Ww, Ff, Gg)

Question #8

Which combination of gametes can these parents produce that would combine to create a curly-winged, fire breathing, dark gray whiptail baby?

○ Dad (w,F,g) and Mom (w,f,g)
○ Dad (W,F,G) and Mom (w,f,g)
○ Dad (w,F,g) and Mom (W,f,g)
○ Dad (w,f,g) and Mom (w,F,G)

Question #9

Which combination of gametes can these parents produce that would combine to create a straight-winged, fire-breathing, light gray whiptail baby?

○ Dad (W,f,g) and Mom (W,f,g)
○ Dad (W,f,g) and Mom (w,F,g)
○ Dad (w,F,g) and Mom (W,f,G)
○ Dad (w,f,g) and Mom (w,F,G)

Figure 2. Example post-test question.

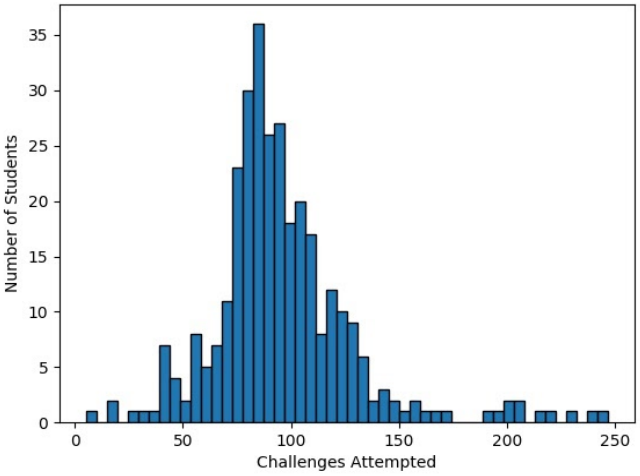


Figure 3. Histogram of students' gameplay trajectories.

Each of the test items is recorded in a binary format, with 1 if the student answered the item correctly and 0 if the item is answered incorrectly. To calculate the competency score for each concept, the total number of correctly answered items was divided by the total number of items for that concept, resulting in a score between 0 and 1. These scores serve as the target labels for our regression models.

5. METHODOLOGY

We evaluate two different approaches to the student competence modeling: single-task and multi-task. The single-task approach involves training an individual model for every concept, with each model only predicting a single competency score. This architecture allows for the model to focus exclusively on modeling trends and correlations between the students' gameplay features and a single competency and does not take into account any interrelationships between the gameplay and multiple concepts. The multi-task approach requires a single model trained to approximate all competency scores using a single 16-unit vector. This approach is advantageous as it is capable of modeling complex, non-linear relationships between the various concept-level competencies that exist within the gameplay data. Multi-task modeling has seen an increase in usage due to its reduced number of parameters to be estimated, as well as the computational time required to train a model for each dependent variable, compared to the single-task modeling technique. Multi-task models' capability to robustly model inherent relationships between multiple dependent variables using a shared input vector space makes this modeling technique ideal for stealth assessment frameworks, as well as circumstances where a large amount of training data may not be readily available [9].

We evaluate these two approaches using two different feature representations of the students' gameplay data: static and sequential representations. The static representation of the data involves producing a single feature vector representative of each student's overall interaction with the *Geniventure* learning environment, resulting in a single dataset of 316 total data samples. The sequential representation is used to model subsequences extracted from individual challenge-level interactions across each student's gameplay trajectory, retaining temporal information based on the order the challenges were completed. This sequence sub-sampling approach results in a single dataset of 29,977 total data samples.

5.1 Interaction Data

Gameplay interactions with the *Geniventure* game environment were recorded in a timestamped log file for each student. The trace data log is a raw event stream in JSON format which records fine-grained information about students' actions in the game, such as a navigated challenge, changed allele, submitted answer, and received hints from the system. The types of actions vary among challenges because of the differences in the challenge settings. To eliminate the influence of the differences in challenges, we defined 10 generic measurements across different challenges that describe contextual information about the challenge itself. The remaining features summarize students' performance and actions within an individual challenge. For each student, we generated his/her gameplay trajectory across each individual challenge attempted. The length of the challenge level trajectories varied from 5 to 248 (M = 95.86, SD = 33.63). Each of the features forms the basis for the static and sequential data.

Table 1. Competency model concepts

Concept Number	Concept Description	Number of Questions
C1	Only one dominant allele is needed to produce the dominant trait.	3
C2	Two recessive alleles are needed to produce a recessive trait.	2
C3	Create or select parental gametes to create an individual offspring with a specific phenotype.	4
C4	Set parental genotypes to produce a specific pattern of offspring.	6
C5	Use patterns in the phenotypes of a group of offspring to predict the genotype of the parents.	5
C6	For some traits primarily influenced by a single gene, both alleles will have some effect, with neither being completely dominant.	2
C7	Breed with a recessive animal to determine an unknown genotype (testcross).	2
C8	Different versions of a gene correspond to (lead to the construction of) different versions of a specific protein.	2
C9	Proteins do work or have jobs to do in cells.	1
C10	Proteins are nanomachines; different proteins do different jobs.	1
C11	The function of a protein is determined by its shape.	1
C12	Different versions of a specific protein have different structures and may also have different functions.	1
C13	Some traits have multiple alleles, which can form a ranked series in terms of dominance.	2
C14	Genes on the X chromosome are referred to as X-linked. Males receive only one copy of the X chromosome and pass on their X only to their daughters.	1
C15	Working from the phenotype, determine possible genotypes for an organism.	2
C16	Use a genotype to predict the phenotype for an organism.	2

The features representing each challenge undertaken by a student are: (1) Pre-test score, (2) level of challenge, (3) mission number of challenge, (4) challenge number, (5) total time spent on challenge, (6) number of movements made during challenge, (7) number of hints encountered during challenge, (8) number of correct movements made during challenge, (9) number of wrong movements made during challenge, and (10) student's completion status of challenge (0: incomplete, 1: complete with wrong answer, 2: complete with correct answer).

5.2 Static Competency Models

We evaluate five different regression models to determine their capabilities to predict students' competency levels for each concept. The features selected for the static competency models summarize the whole gameplay of each student across all challenges and levels. Using the challenge-level features noted above, the summative student-level features generated for the static models are (1) average time spent per challenge, (2) total time spent playing challenges, (3) fraction of challenges failed, (4) fraction of challenges succeeded, (5) fraction of challenges abandoned, (6) fraction of incorrect movements, (7) fraction of correct movements, (8) total hints received, (9) number of hints per level, (10) hint count per challenge, and (11) number of levels played.

We evaluate two variations of static modeling techniques: single-task and multi-task. Single-task models predict each target concept score as an independent regression problem. The data set and features used in each model are identical, but the target variable is a single competency score for each model. Multi-task models approximate all target variables in a single model. However, not all of the static, single-task models can effectively translate to a multi-task learning environment. Using single-task learning, we aim to discover the best model for each target variable independently while multi-task models perform better when there are underlying dependencies between the various competencies and a student's gameplay features.

5.2.1 Single-Task Models

We evaluate three single-task models. Elastic Net is a linear regression model that utilizes both L1 and L2 regularizations. The hyperparameter tuning of Elastic-Net was performed on the L1 and L2 regularization coefficients (alpha, L1 ratio). Gradient-Boosted Regression (GBR) is a decision tree-based modeling approach that builds an ensemble of weak predictors to approximate the target variable. The model is built in an iterative fashion where each subsequent stage improves on the model created in the previous stage. The hyperparameter tuning for the GBR model was based on fine-tuning the maximum depth of each tree in the model and the total number of estimators added to the model. We also evaluate a Random Forest regressor, another type of ensemble learning method using a 'forest' of decision trees. Each tree is randomly assigned insensitivity to different features in the training data (i.e., feature bagging). This approach allows for larger model ensembles while avoiding overfitting. The hyperparameter tuning for Random Forest was performed on the maximum depth of each tree and the total number of trees in the forest. While both Random Forest and Gradient Boosted Trees are decision tree-based ensemble learners, a notable difference between these two models is how the trees are added to the ensemble. Within Random Forest models, trees are added independently while in GBR models, trees are added incrementally to compensate for the shortcomings of the previous iteration of models. For the single-task approach, we use a single model for each competency score. We keep the regression model type consistent across all competencies and the hyperparameter values consistent across models.

5.2.2 Multi-Task Models

Due to constraints in models, only certain types of algorithms support multi-task modeling. In this work, two types of multi-task regression models are tested: the multi-task version of Elastic Net and the multi-task version of Random Forest. The multi-task version of Elastic Net adds the constraint that the selected features in the model are the same for all the tasks. The Random Forest regressor is one of the few models that does not require any special modification to support multi-task learning due to the trees in the regressor being built on different subsamples of the dataset.

Each of the static models was implemented using the scikit-learn library in Python. The data set (316 samples) is divided randomly into an 80/20 split with 20% serving as a held-out test set to evaluate the models. The 80% split is used for training the models, with five-fold cross-validation being applied to determine the best model. The training and test splits remain consistent across all investigated models and configurations (e.g., static vs. sequential, single-task vs. multi-task) to ensure a fair comparison between models, as well as the five-fold cross-validation splits. The final hyperparameter values as a result of the cross-validation on the training data are shown in Table 2.

Table 2. Static model hyperparameters

Regression Model	Task Type	Hyperparameters
Elastic Net	Single-Task	alpha = 0.05 L1 Ratio = 0.9
Gradient Boosted Regression	Single-Task	Max Tree Depth = 2 Number of Trees = 20
Random Forest	Single-Task	Max Tree Depth = 3 Number of Trees = 250
Elastic Net	Multi-Task	alpha = 0.2 L1 Ratio = 0.9
Random Forest Regression	Multi-Task	Max Tree Depth = 2, Number of Trees = 200

5.3 Sequential Competency Models

We explore four different types of deep learning-based models to model sequential representations of each student's gameplay information across attempted challenges. Here, the motivation is to determine whether providing sequential context for each of the student's problem-solving behaviors induces higher performance when modeling the competencies. To provide further sequential information to each of the models, we generate additional temporal features averaged across all challenges completed up to the current challenge attempted by the user: (1) average time per challenge, (2) average movements per challenge, (3) average correct movements per challenge, (4) average incorrect movements per challenge, (5) average hint count per challenge, (6) average unsubmitted challenges, (7) average failed challenges, and (8) average successful challenges.

We use these features in addition to the 10 static challenge-level features described in Section 5.1 to provide a total of 18 features to each of the sequential models. The models used for both the single-task and multi-task sequential models are variants of recurrent neural networks including Long Short-Term Memory recurrent neural networks (LSTMs) [11] and Gated Recurrent Units (GRUs) [6], due to their capability to model both single-task and multi-task data. LSTMs utilize a sequence of memory blocks that each contain an input gate, forget gate, and an output gate. The forget gate determines whether the previous memory block's gradient is retained or discarded, thus allowing the LSTM to model long-term dependencies across temporal sequences, while the input and output gates modulate the input and output vectors, respectively. GRUs are mechanisms that provide the same "forgetting" functionality as LSTMs but contain fewer hyperparameters, utilizing an update gate and a reset gate. This allows GRUs to be more computationally efficient and sometimes more effective on less training data than LSTMs.

In addition to standard LSTMs, we evaluate bidirectional LSTMs (Bi-LSTMs) [32] as well as LSTMs implementing a self-attention mechanism (SA-LSTMs) [41]. Bidirectional LSTMs are a variation of LSTMs that contain two input layers on opposing sides of the hidden layer, allowing the model to retain temporal information based on the past and the future of the input sequence, as opposed to only the past. A self-attention LSTM provides additional temporal context beyond contiguous feature vectors by utilizing a weighted sum of hidden representations of the entire sequence.

Adopting the same manner used in training the static competency models, each model is optimized using 5-fold cross validation, where the data splits are consistent across both static and sequential models to ensure a fair comparison, and then evaluated with the held-out test set. The hyperparameters are tuned using an iterative grid search, and each model was trained for 200 epochs. The subsequences used to train each sequential model were sampled across the challenges completed by each student using a sequence length of 10, and a sampling stride of 1. We use front padding in each sequence during the subsampling process to allow the models to fit during the beginning of each sequence. The concept-level prediction made for each student was calculated by taking the average competency prediction value across an entire sequence. The sequential data modeling pipeline was implemented using Python, and the deep learning models were implemented using the Keras library with the TensorFlow backend. The hyperparameter tuning was performed across the number of hidden units in each model's hidden layer, as well as the dropout rate in the hidden layer [9]. The final hyperparameter values as a result of the cross-validation on the training data are shown in Table 3.

5.3.1 Single-Task Models

To evaluate the single-task sequential modeling approach, we train 16 different independent models, with each model approximating a single competency score based on the gameplay features described in Sections 5.1 and 5.3. Using the cross-validation performance on the training data, we selected the optimal configuration for each model type based on the highest performance in terms of the average R^2 value across all competency scores.

5.3.2 Multi-Task Models

Because of the architecture of the sequential deep learning models, each single-task model type is also able to perform as a multi-task model, with the only change occurring within the output layer, as the number of output units is expanded to contain an individual output node for each concept, instead of a single concept. This allows a single model to simultaneously infer student competencies across all concepts. Similar to the single-task models, the optimal model configurations were selected based on the average R^2 score across all concepts.

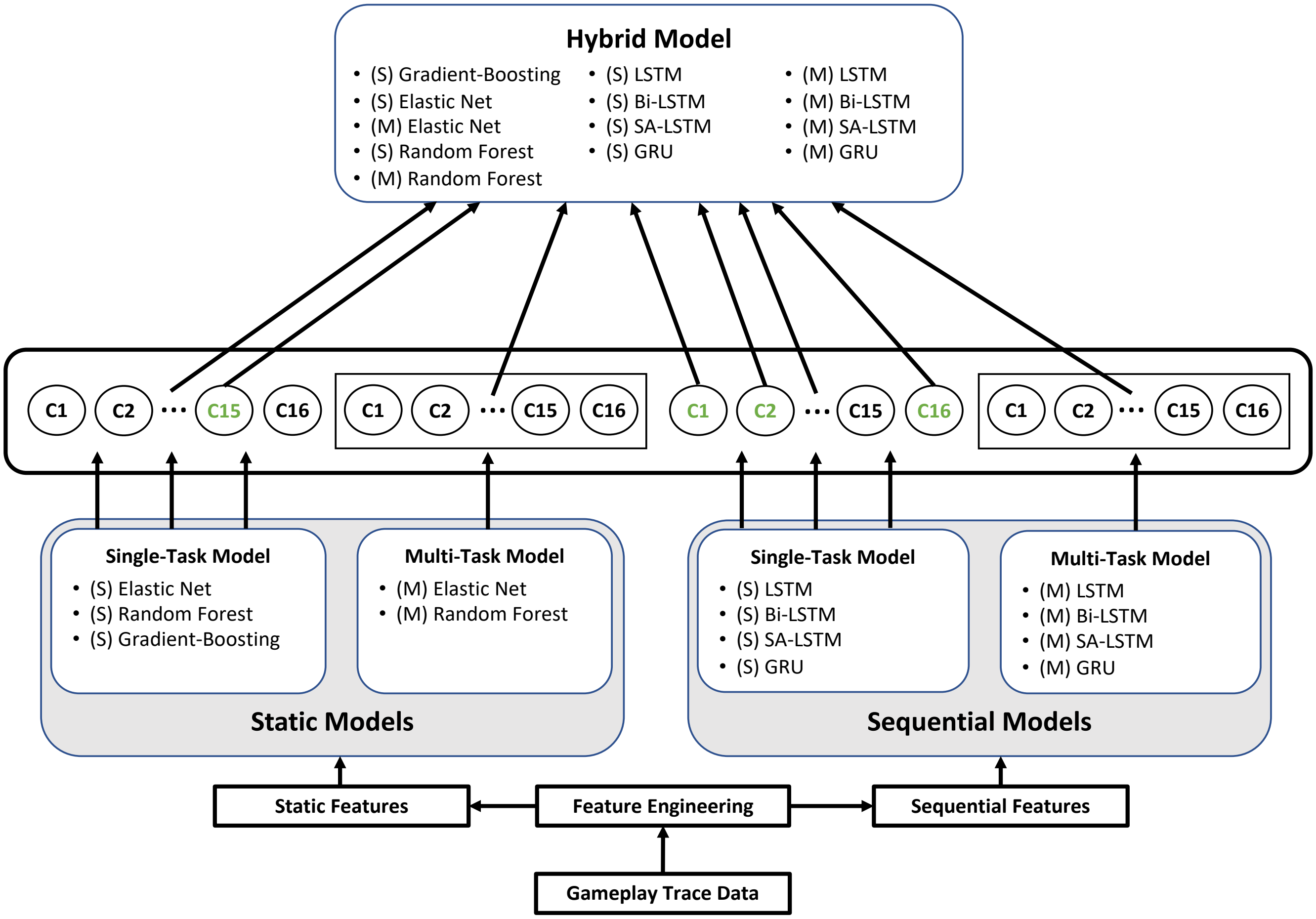


Figure 4. Hybrid, static, and sequential stealth assessment models.

We hypothesize that the optimal modeling techniques will vary due to the different complex characteristics that underlie each concept and the varying fitness of ECD models to these characteristics. Therefore, we propose the use of a hybrid framework that contains a combination of the various single-task and multi-task models that are both static and sequential. A visualization of the proposed hybrid stealth assessment framework comprised of the optimal models (highlighted in green text) is shown in Figure 4 above.

The rationale behind the use of a hybrid stealth assessment framework is that student competencies can vary widely with regards to each concept's correlation to specific questions in post-test scoring methods, as well as each concept's correlation to specific gameplay features or levels. By implementing both static and sequential variations of single-task and multi-task models, the long-term and short-term tendencies within each student's gameplay is explored on a challenge and a student level. Additionally, the relationships between the individual competencies are modeled independently in the single-task approach, indicating whether certain concepts have no interweaving tendencies with other concepts within the gameplay. By utilizing a mixture of both single-task and multi-task models in this framework, multi-task models are only fit where underlying relationships exist between concepts, and concepts that have no underlying relationships with other concepts are optimally modeled by the single-task approach. The same concept applies to the sequential and static modeling: only concepts that have informative temporal trends across a student's challenge-level gameplay data are modeled by the sequential models. All other concepts are modeled by the static models utilizing only student-level data.

Table 3. Sequential model hyperparameters

Regression Model	Task Type	Hyperparameters
LSTM	Single-Task	Hidden units = 80 Dropout rate = 0.33
Bi-directional LSTM	Single-Task	Hidden units = 20 Dropout rate = 0.33
GRU	Single-Task	Hidden units = 80 Dropout rate = 0.5
Self-attention LSTM	Single-Task	Hidden units = 60 Dropout rate = 0.33
LSTM	Multi-Task	Hidden units = 100 Dropout rate = 0.5
Bi-directional LSTM	Multi-Task	Hidden units = 60 Dropout rate = 0.33
GRU	Multi-Task	Hidden units = 40 Dropout rate = 0.33
Self-attention LSTM	Multi-Task	Hidden units = 80 Dropout rate = 0.5

Table 4. R^2 value of single-task models based on held-out test set

Model	Concept C1	C2	C3	C4	C5	C6	C7	C8	C9	C11	C12	C13	C14	C15	C16	Mean
Elastic Net	0.210	0.169	0.176	0.300	0.348	0.232	0.145	0.185	0.073	0.138	0.048	0.098	0.052	0.163	0.094	0.162
GBR	0.214	0.301	0.156	0.363	0.365	0.119	0.107	0.190	0.045	0.111	0.034	0.110	0.081	0.263	0.100	0.171
RF	0.269	0.297	0.165	**0.434**	**0.404**	0.109	0.165	**0.241**	**0.096**	**0.225**	0.029	0.086	**0.100**	**0.318**	0.148	**0.206**
LSTM	0.314	**0.383**	0.149	0.398	0.346	0.155	0.141	0.157	0.054	0.091	0.034	0.013	0.028	0.302	0.185	0.183
Bi-LSTM	**0.363**	0.328	0.164	0.376	0.368	0.153	0.072	0.104	-0.029	0.075	0.009	-0.020	-0.094	0.273	**0.262**	0.160
SA-LSTM	0.315	0.351	0.135	0.356	0.314	0.148	0.107	0.123	0.009	0.070	0.050	0.038	-0.030	0.306	0.218	0.167
GRU	0.109	0.088	0.062	0.156	0.189	0.090	0.092	0.029	-0.158	0.017	0.022	0.004	-0.319	0.089	0.031	0.033

Table 5. R^2 value of multi-task models based on held-out test set

Model	Concept C1	C2	C3	C4	C5	C6	C7	C8	C9	C11	C12	C13	C14	C15	C16	Mean
RF	0.279	0.298	0.131	0.336	0.337	0.154	0.152	0.170	0.063	0.129	0.051	**0.114**	0.051	0.295	0.193	0.184
Elastic Net	0.211	0.171	**0.182**	0.307	0.350	**0.239**	0.159	0.181	0.077	0.136	0.057	0.099	0.049	0.172	0.080	0.165
LSTM	0.291	0.270	0.144	0.362	0.346	0.147	0.179	0.174	0.075	0.131	**0.058**	0.024	0.029	0.259	0.130	0.175
Bi-LSTM	0.313	0.273	0.157	0.371	0.356	0.142	0.176	0.166	0.066	0.123	**0.058**	0.021	0.014	0.260	0.144	0.176
SA-LSTM	0.320	0.302	0.176	0.361	0.352	0.173	**0.201**	0.133	0.023	0.110	0.017	0.034	-0.055	0.313	0.255	0.181
GRU	0.309	0.241	0.152	0.352	0.350	0.156	0.199	0.169	0.061	0.127	0.050	0.048	0.036	0.275	0.040	0.171

Table 6. Highest R^2 values of optimal hybrid competency models

Model	Concept C1	C2	C3	C4	C5	C6	C7	C8	C9	C11	C12	C13	C14	C15	C16	Mean
Hybrid	0.363	0.383	0.182	0.434	0.404	0.239	0.201	0.241	0.096	0.225	0.058	0.114	0.100	0.318	0.262	**0.241**

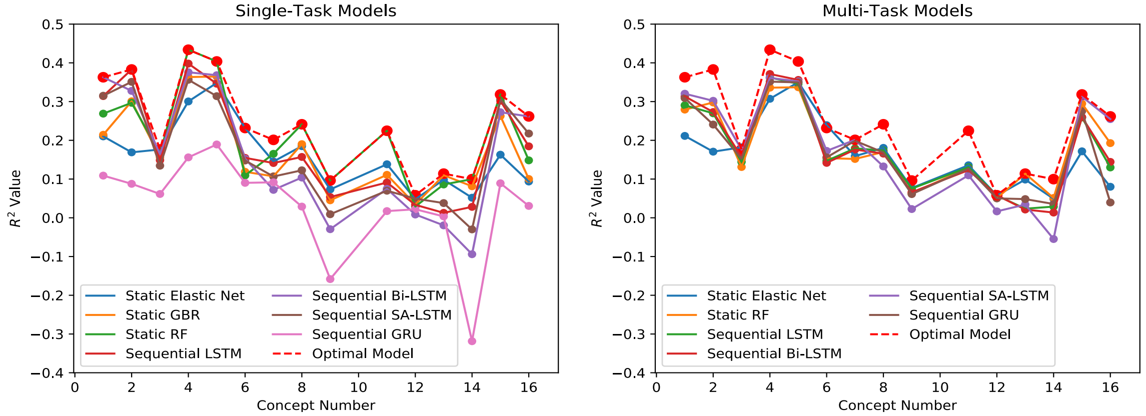


Figure 5. Performance of single-task and multi-task models compared to hybrid model performance.

6. RESULTS AND DISCUSSION

We report the results of the single-task models (Table 4) and the multi-task models (Table 5) for each concept in terms of R^2. The highest R^2 value produced for each individual concept is presented in Table 6, as this represents the performance of our proposed hybrid framework across all concepts. Figure 5 shows the performance of single-task and multi-task models compared to the performance of the optimal hybrid model (Left: comparison to single-task models, Right: comparison to multi-task models). The results are obtained based on each model's performance on the held-out test set after being trained on the entirety of the training set. As noted above, the cross-validation splits applied to the training set were performed on a student level to prevent data leakage and were consistently applied to the set of machine learning techniques for a fair comparison across different models. For this work, Concept 10 was omitted because every student that took the post-test survey answered the question correctly, resulting in a dependent variable with zero variance, thus having no impact on the evaluation of our respective models.

The best performing model in terms of average R^2 value across all concepts was the single-task Random Forest; however, it was only the optimal model for 7 out of the 15 total concepts. The single-task bidirectional long short-term memory network performed the highest for two concepts, as well as the multi-task Elastic Net. The single-task LSTM and the multi-task Random Forest (RF), Bi-LSTM, and self-attention LSTM were optimal models for one concept each. The Gradient Boosted Regression (GBR), single-task

Gated Recurrent Unit, and single-task SA-LSTM performed relatively poorly and were not the highest performing competency models for any of the concepts. The multi-task models were the most effective approach for five of the 15 concepts, while single-task models were most effective for the other 10 concepts.

Across both single-task and multi-task models, the GRU was the lowest-performing model, achieving an average R^2 value of 0.102 across all concepts. The results that variants of LSTMs (e.g., standard LSTMs, SA-LSTMs, Bi-LSTMs) achieved the highest R^2 score in predicting student competencies on at least one concept demonstrate that there exist complex, sequential patterns within the students' gameplay data, which were effectively modeled by the LSTMs' three gating units, but not by the two gating units enabled in GRUs. The SA-LSTMs, Bi-LSTMs, and standard LSTM models all returned relatively equal performances across the single-task and multi-task data, with average R^2 values of 0.174, 0.168, and 0.179, respectively. It appears that although the Bi-LSTM and SA-LSTM capture various extra temporal contextual patterns not inherently captured by the standard LSTM, this information is not globally beneficial to all the competency models, explaining why neither model outperforms the standard LSTM on average. However, this result might also be attributed to the fact that the sequential data was only generated from 316 students, which may not be enough information for any of the more complex, sequential models utilizing a higher number of trainable parameters, to truly detect informative underlying temporal patterns.

Selecting the single-task RF as the model for all concepts based on its average performance across all the concepts results in a mean R^2 value of 0.206. However, as illustrated in Table 6, by using our proposed hybrid system approach and selecting the optimal model for each individual concept, we can obtain a performance of 0.241, which is a 17.0% improvement compared to a homogenous framework typically used within stealth assessment. Our observation that the use of multiple models in the hybrid stealth assessment framework would induce higher performance than using a single model can be explained by the fact that static, sequential, single-task, and multi-task models were all selected as an optimal model at least once.

Additionally, it should be noted that when considering only the concepts that mapped to multiple questions (i.e. 1-8, 13, 15-16), the deep-learning based sequential models produced a higher and more consistent performance (0.224 for single-task, 0.222 for multi-task) than the static models (0.214 for single-task, 0.210 for multi-task) on average. The multi-task SA-LSTM and the single-task RF both achieved the optimal performance across the multi-question concepts, with an average R^2 value of 0.239. Random Forest may also perform relatively well as a competence model because it uses an ensemble approach, making it more robust against overfitting.

One correlation that was noted is that the single-task models were the best technique for 75% (3 out of 4) of the concepts that had only one corresponding question in the post-test. This can potentially be attributed to the fact that each of the single-task models only models a single concept, without taking into account any of the linear and non-linear relationships that might exist between the gameplay features and the different competencies for a single student. Concepts that correspond to only a single question possibly contain a less complex relationship between the competency scores and the gameplay features, meaning that a single-task model is sufficient for that modeling task without simultaneously modeling any context related to competencies for other concepts, which can have a detrimental impact to the predictive tasks. In addition, each of the three optimal models for the single-question concepts were trained using static feature representations, suggesting that the student-level features were the most informative to our model, and the temporal information did not yield greater predictive performance for the student competency models.

However, we also observe that the single-task models were also frequently the highest-performing models for the multi-question concepts. Seven out of the 11 concepts that were represented by multiple post-test questions were optimally modeled using a single-task model, either using static or sequential representations. Out of these 7 highest-performing models, 4 of them used static input representations. In a similar manner to the single-task models mentioned previously, this implies that student-level features were informative for a subset of the multi-question concepts, while the temporal context provided within the sequence modeling tasks was still beneficial to predicting students' individual competencies for the three other concepts.

Overall, the majority of optimal classifiers across the single-question and multi-question concepts were single-task, static representations, as these account for 7 out of the 15 total concepts we evaluated. We then analyze the remaining models to investigate if there are any correlations between the concepts and the optimal models. The competency models for Concepts 1 and 2 were both modeled using sequential single-task models, two concepts that correspond to five combined post-test questions. Concept 1 deals with generating dominant traits using alleles, while Concept 2 deals with a similar task generating recessive traits using alleles. The similarity in these two concepts may be a possible reason that the highest predictive performance was achieved by the same modeling approach. The highest R^2 values (0.434 and 0.404) occurred in Concepts 4 and 5, which are the two concepts that correspond to 6 and 5 post-test questions, respectively. The correlation between the higher performance in these two RF-based competency models can be explained by the fact that ensemble models leveraging more single-task models contribute to improvement of the average predictive performance, which prevents a model that produces a less accurate prediction from heavily impacting the overall representative performance.

The relative scores between concepts are highly correlated across modeling methods. In other words, the concepts that had a high R^2 score for one model also had a high R^2 score for most of the other models. As shown in Figure 5, Concepts 4 and 5 have the highest R^2 value irrespective of the modeling method, and Concept 14 is on the lower end of R^2 values. This could be because of how well a gameplay feature predicts a concept is dependent on the type of concept. In other words, some concepts are harder to model irrespective of the modeling approach used for the model.

Interestingly, concepts that contained only a single question (i.e. 9, 11-12, 14) produced noticeably low R^2 values. These single-question concepts produced an average R^2 value of 0.046 across all the models. Because there was only a single question associated with the concept, each competency score was entirely dependent on students' single response to the question, which could result in a reliability issue in the competency scores due to students' behaviors related to guess and slip as well as a higher variance in the scores, together possibly attributing to these low R^2 scores.

A chart of the average student score for each concept based on their post-test performance is shown in Figure 6 below, distinguishing between single-question and multi-question concepts. The average student performance on multi-question concepts was markedly higher than for single-question concepts, with students achieving scores of 0.672 and 0.507, respectively. It was noted that the average students' scores on the questions mapped to a single concept was remarkably low compared to an overall student score of 0.661 across all questions. This factor may also have impacted the predictive performance of the competency models compared to competency models that encounter fairly consistent or accurate student answers to post-test questions, as the questions corresponding to lower student scores introduce higher variance into the resulting competency scores used to train the models.

Finally, we investigate the impact that overlapping concepts may have on the performance of the classifiers. A concept is considered to be "overlapping" if it shares a correlated question with one or more different concepts. Out of the 16 concepts, 6 concepts were found to be "completely" overlapping; that is, every question associated with that concept was also associated with another concept. One concept was "partially" overlapping, indicating that only a portion of the associated questions were also mapped to another concept. The remaining concepts were the only ones that corresponded with their own associated question or group of questions. Student scores were significantly higher for overlapping concepts as opposed to non-overlapping questions, achieving average scores of 0.718 and 0.550, respectively. This trend is also present in our competency models, as the optimal models in our hybrid framework yield R^2 values of 0.339 (overlapping) and 0.157 (non-overlapping) on average. Surprisingly, the optimal models for the overlapping concepts were primarily single-task models, with the exception of one model. This indicates that multi-task modeling across all the concepts including both relevant and irrelevant concepts is actually detrimental in terms of achieving higher predictive performance. Thus, a promising future direction is to investigate multi-task learning performance by grouping relevant concepts and separately modeling related concepts only.

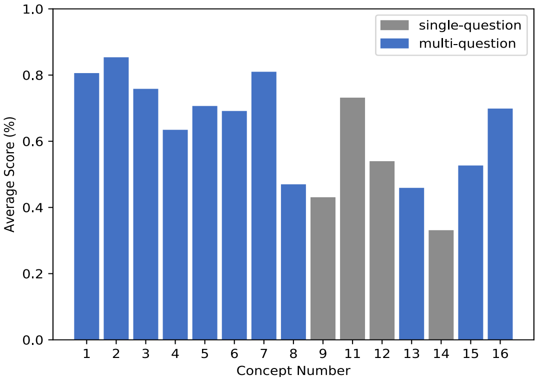


Figure 6. Student post-test performance on concepts associated with single and multiple questions.

In this particular application domain (genetics), concepts C1 and C2 are foundational to eight other concepts, as they describe a common pattern of gene variant behavior in inheritance of traits. Some concepts are related variously to other concepts, e.g., C5 requires deductive reasoning based on C1, C2, and C15, while it also serves as a prerequisite for C7, which allows determining gene variants for ambiguous traits. Alternatively, concepts C9-C12 focus primarily on molecular genetic inheritance and are not as tightly related to other concepts. This example of varying connections within genetics-related concepts illustrates the broader application of our hybrid model and why it demonstrates promise for other domains.

7. CONCLUSION

Stealth assessment holds considerable potential for game-based learning. Recent work exploring stealth assessment has typically employed a single machine learning technique to devise competency and evidence models. This approach operates under the assumption that each student competency can be optimally modeled by the same learning algorithm that yields the highest predictive performance on average. However, this may not always be the case, as student competencies often have varying interleaving relationships with each other or even underlying complexities within itself.

In this work, we demonstrate the effectiveness of a hybrid stealth assessment framework consisting of a combination of single-task and multi-task models, using static and sequential features to represent student gameplay data. We evaluate our stealth assessment framework using a game-based learning environment and predict student competencies as measured by a post-test. Results indicate that a heterogeneous approach to stealth assessment modeling techniques induces higher results across all concepts when compared to the single-model baseline evaluations. Selecting a single competency model for all concepts based on its average performance across all the concepts is a common practice in stealth assessment frameworks. However, the proposed hybrid system using the optimal model for each individual concept returns a performance that is substantially higher than a homogeneous framework. In addition to static, single-task modeling, the sequential, multi-task modeling approach can adapt to multiple concepts by effectively capturing sequential context underlying individual students' gameplay behaviors, as well as simultaneously modeling various competencies that were manifested throughout the gameplay sessions. The use of all of the aforementioned modeling techniques provides a multi-dimensional approach that has been demonstrated to be a step forward in improving stealth assessment techniques.

There are a number of future directions that can be investigated to further improve the performance of the hybrid stealth assessment framework. Multi-task learning becomes increasingly difficult as the number of tasks increases and training deep sequential models for 16 tasks using only 316 data samples is likely a limiting factor in the multi-task models' performances. To gain further insight into the use of multi-task learning as a competency modeling technique, the hybrid stealth assessment framework presented in this work should be evaluated on comparatively larger datasets. This also enables the evaluation of the hybrid framework's ability to adequately translate to other student populations. Alternatively, different ways to reduce the number of tasks can be investigated. Due to the hierarchical, interweaving relationships within both individual concepts and between concepts and various questions, it will be worthwhile to investigate other sophisticated hierarchical modeling methods such as Bayesian hierarchical modeling or clustering methods, as well as refine the post-test questions and the mapping to the concepts to more reliably assess students' competency for each concept. Additionally, the feature engineering process performed for both static and sequential models can evolve significantly, possibly inducing higher performance from the competency models. Finally, it will be instructive to investigate the generalizability of this framework across different learning environments, contexts, and student populations.

8. ACKNOWLEDGEMENTS

The authors would like to thank Robert Taylor for his assistance in facilitating this research. This research was supported by the National Science Foundation under Grant DRL-1503311. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

9. REFERENCES

[1] 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 the 11th International Conference on Educational Data Mining*. International Educational Data Mining Society, 208-218.

[2] Asbell-Clarke, J., Rowe, E., Sylvan, E., and Baker, R. 2013. Working through impulse: assessment of emergent learning in a physics game. *Games+ Learning+ Society*. 9.

[3] Bosch, N., Chen, H., D'Mello, S., Baker, R. and Shute, V. 2015. Accuracy vs. availability heuristic in multimodal affect detection in the wild. In *Proceedings of the 2015 International Conference on Multimodal Interaction*. 267-274.

[4] Buffum, P. S., Frankosky, M., Boyer, K. E., Wiebe, E. N., Mott, B. W., and Lester, J. C. 2016. Collaboration and gender equity in game-based learning for middle school computer science. *Computing in Science and Engineering.* 18, 2, 18-28.

[5] Chaudhry, R., Singh, H., Dogga, P., and Saini, S. K. 2018. Modeling hint-taking behavior and knowledge state of students with multi-task learning. In *Proceedings of the International Conference on Educational Data Mining.* International Educational Data Mining Society.

[6] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078.*

[7] 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.

[8] Falakmasir, M.H., Gonzalez-Brenes, J.P., Gordon, G.J. and DiCerbo, K.E. 2016, April. A data-driven approach for inferring student proficiency from game activity logs. In *Proceedings of the Third ACM Conference on Learning@ Scale.* 341-349.

[9] Geden, M., Emerson, A., Rowe, J., Azevedo, R., and Lester., J. 2020 (in press). Predictive student modeling in educational games with multi-task learning. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence.*

[10] Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., and Edwards, T. 2016. Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior.* 54, 170-179.

[11] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*. 9, 8, 1735-1780.

[12] Karumbaiah, S., Baker, R. S., and Shute, V. 2018. Predicting quitting in students playing a learning game. In *Proceedings of the 11th International Conference on Educational Data Mining.* 167-176.

[13] Kiili, K., Devlin, K., Perttula, T., Tuomi, P., and Lindstedt, A. 2015. Using video games to combine learning and assessment in mathematics education. *International Journal of Serious Games*. 2, 4, 37-55.

[14] Kim, Y.J., Almond, R.G. and 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.

[15] Lester, J.C., Ha, E.Y., Lee, S.Y., Mott, B.W., Rowe, J.P. and Sabourin, J.L. 2013. Serious games get smart: Intelligent game-based learning environments. *AI Magazine.* 34, 4, 31-45.

[16] Liu, Z., Zhi, R., Hicks, A., and Barnes, T. 2017. Understanding problem solving behavior of 6–8 graders in a debugging game. *Computer Science Education*. 27, 1, 1-29.

[17] Ma, Y., Cui, C., Yu, J., Guo, J., Yang, G., and Yin, Y. 2019. Multi-task MIML learning for pre-course student performance prediction. *Frontiers of Computer Science*. 14, 5, 145313.

[18] McElroy-Brown, K. and Reichsman, F. 2019. Genetics with dragons: Using an online learning environment to help students achieve a multilevel understanding of genetics. Retrieved from http://concord.org.

[19] Mills, C., D'Mello, S., Lehman, B., Bosch, N., Strain, A., and Graesser, A. 2013. What makes learning fun? exploring the influence of choice and difficulty on mind wandering and engagement during learning. In *Proceedings of the International Conference on Artificial Intelligence in Education*. Springer, 71-80.

[20] Min, W., Frankosky, M., Mott, B.W., Rowe, J., Smith, P.A.M., Wiebe, E., Boyer, K., and Lester, J. 2019. DeepStealth: game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies.*

[21] Min, W., Frankosky, M. H., Mott, B. W., Wiebe, E. N., Boyer, K. E., and Lester, J. C. 2017. Inducing stealth assessors from game interaction data. In *Proceedings of the 9th International Conference on Artificial Intelligence in Education.* Springer, 212-223.

[22] Min, W., Frankosky, M.H., Mott, B.W., Rowe, J.P., Wiebe, E., Boyer, K.E. and Lester, J.C. 2015. DeepStealth: leveraging deep learning models for stealth assessment in game-based learning environments. In *Proceedings of the International Conference on Artificial Intelligence in Education.* Springer, Cham, 277-286.

[23] Min, W., Mott, B. W., Rowe, J. P., Liu, B., and Lester, J. C. 2016. Player goal recognition in open-world digital games with long short-term memory networks. In *Proceedings of the International Joint Conference on Artificial Intelligence.* 2590-2596.

[24] Mislevy, R., Steinberg, L., and Almond R. 2003. Focus article: on the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives.* 1, 1, 3–62.

[25] Owen, V. E., Roy, M. H., Thai, K. P., Burnett, V., Jacobs, D., Keylor, E., and Baker, R. S. 2019. Detecting wheel-spinning and productive persistence in educational games. International Educational Data Mining Society.

[26] Puntambekar, S. and Hubscher, R. 2005. Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed?. *Educational Psychologist.* 40, 1, 1-12.

[27] Qu, S., Li, K., Wu, B., Zhang, X., and Zhu, K. 2019. Predicting student performance and deficiency in mastering knowledge points in MOOCs using multi-task learning. *Entropy.* 21, 12, 1216.

[28] Roschelle, J., Dimitriadis, Y. and Hoppe, U. 2013. Classroom orchestration: synthesis. *Computers & Education*. 69, 523-526.

[29] Rosenheck, L., Lin, C., Klopfer, E., and Cheng., M. 2017. Analyzing gameplay data to inform feedback loops in the radix endeavor. *Computers & Education*. 111, 60–73.

[30] Rowe, J. P., Shores, L. R., Mott, B. W., and Lester, J. C. 2011. Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education.* 21, 1-2, 115-133.

[31] Sawyer, R., Smith, A., Rowe, J., Azevedo, R. and Lester, J. 2017. Enhancing student models in game-based learning with facial expression recognition. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization.* 192-201.

[32] Schuster, M. and Paliwal, K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing.* 45, 2673 - 2681.

[33] Shute, V. J. and Ke, F. 2012. Games, learning, and assessment. *Assessment in Game-Based Learning*. Springer, 43-58.

[34] Shute, V. J. 2011. Stealth assessment in computer-based games to support learning*. Computer Games and Instruction.* 55, 2, 503-524.

[35] Shute, V., Ventura, M., Zapata-Rivera, D., and Bauer, M. 2009. Melding the power of serious games and embedded assessment to monitor and foster learning flow and grow. *Serious Games: Mechanisms and Effects*. 2, 295–321.

[36] Shute, V. and Ventura, M. 2013*. Measuring and Supporting Learning in Games: Stealth Assessment.* The MIT Press, Cambridge, MA.

[37] Shute, V.J. and Moore, G.R. 2017. Consistency and validity in game-based stealth assessment. *Technology enhanced innovative assessment: Development, Modeling, and Scoring from an Interdisciplinary Perspective.* 296.

[38] States, N. L. 2013. Next Generation Science Standards. Washington.

[39] Tabak, I. 2004. Synergy: A complement to emerging patterns of distributed scaffolding. *The Journal of the Learning Sciences*. 13, 3, 305-335.

[40] Tokac, U., Novak, E., and Thompson, C. G. 2019. Effects of game-based learning on students' mathematics achievement: A meta-analysis. *Journal of Computer Assisted Learning*. 35, 3, 407-420.

[41] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998-6008.

[42] Wang, P., Rowe, J.P., Min, W., Mott, B.W. and Lester, J.C. 2017. Interactive narrative personalization with deep reinforcement learning. In *Proceedings of the International Joint Conference on Artificial Intelligence.* 3852-3858.

[43] Wouters, P., Van Nimwegen, C., Van Oostendorp, H., and Van Der Spek, E. D. 1993. A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*. 105, 2, 249-267.

[44] Wouters, P. and Van Oostendorp, H. 2013. A meta-analytic review of the role of instructional support in game-based learning. *Computers & Education.* 60, 1, 412-425.