Mars, Minecraft, and AI: A Deep Learning Approach to Improve Learning by Building

Samuel Hum(✉), Evan Shipley(✉), Matt Gadbury(✉), H Chad Lane(✉),
and Jeffrey Ginger(✉)

University of Illinois at Urbana-Champaign, Urbana, IL 61820, USA
{hum3,evanjs3,gadbury2,hclane,ginger}@illinois.edu

Abstract. Middle school students learned about astronomy and STEM concepts while exploring *Minecraft* simulations of hypothetical Earths and exoplanets. Small groups ($n = 24$) were tasked with building feasible habitats on Mars. In this paper, we present a scoring scheme for habitat assessment that was used to build novel multi/mixed-input AI models. Using Spearman's rank correlations, we found that our scoring scheme was reliable with regards to team size and face-to-face instruction time and validated with self-explanation scores. We took an exploratory approach to analyzing image and block data to compare seven different input conditions. Using one-way ANOVAs, we found that the means of the conditions were not equal for accuracy, precision, recall, and F1 metrics. A post hoc Tukey HSD test found that models built using images only were statistically significantly worse than conditions that used block data on the metrics. We also report the results of optimized models using block only data on additional Mars bases ($n = 57$).

Keywords: Scoring Scheme · Artificial Intelligence · Habitat Building · Informal Learning · Minecraft

1 Introduction

Minecraft has become one of the most popular games in the world, with over 140 million monthly users, and 21.21% of daily user traffic originating from the U.S. [7]. Researchers have shown that even with little experience learners master controls quickly and can effectively engage content when *Minecraft* is used in STEM learning environments [3]. Given the ubiquity of the game, as well as ample opportunities to conduct research on learning and motivation [3], *Minecraft* deserves attention regarding effectiveness for promoting positive student outcomes in formal and informal learning environments.

The data used in this study comes from the What-if Hypothetical Implementations in Minecraft (WHIMC) project that uses *Minecraft* as a vehicle for understanding student interest and motivation in exploring STEM content, primarily Astronomy and Earth Science. Learners are presented with a variety of

A. M. Olney et al. (Eds.): AIED 2024, LNAI 14830, pp. 422–430, 2024.
https://doi.org/10.1007/978-3-031-64299-9_39

"what-if" scenarios, such as "What if Earth had a colder sun?". Working through these counterfactual examples of phenomena in science has shown promise of enhancing learning above and beyond studying strictly factual information [4]. A major highlight of the camp experience is the collaborative build phase, where students work with peers to construct habitats on Mars that scientists or explorers might inhabit. A *Minecraft* habitat or base is an assigned region, sized under a few hundred blocks in any dimension, where participants work in teams to build. This paper proposes a novel scoring scheme for work products such as the Mars habitats in *Minecraft* and a novel method for work product assessment in educational video games. We propose the use of multi/mixed-input models in *Minecraft* that takes data from in-game images and materials groups used for their bases to assess learning. Although multimodal data has been used in educational research and video game environments, AI models using different forms of data have not been applied to student work products [6]. Thus, our paper is guided by the following questions and hypotheses:

RQ1. What criteria constitute a reliable scoring scheme for Mars habitat builds in *Minecraft*, and to what extent does the scoring scheme hold across differences in group sizes and amount of face-to-face time with instructors? H1: A comprehensive scoring scheme based on inclusion of essential aspects of habitability will contribute to a reliable scoring scheme. Group size and face-to-face time with instructors will not contribute to significant differences between scores.

RQ2. What, if any, relationship do Mars habitat scores have with learning outcomes? H2: Higher scores on Mars habitats will positively correlate with knowledge assessment scores.

RQ3. What type or combination of data input accessible in Minecraft should be used for artificially intelligent detectors of habitat quality? H3: Incorporating models built using a combination of multiple forms of data concatenated together will outperform models solely built using one form of data. These models can extrapolate information from different aspects of the bases for more accurate predictions that would be impossible for models only built using a single type of data.

2 Background

In this study, we are interested in analyzing what are called, "student work products", referring to specific, task-driven designs and creations, also called "artifacts". In the end of a learning activity, a student has created a product manifesting their conceptual understanding of the content they interacted with throughout a learning experience. Work products as means of assessing learner knowledge and creativity emerged from Constructionism and the idea that knowledge is produced through students' creative and collaborative work [2]. This notion has been fully embraced by the Maker Movement and the desired approach to better understand what tools and activities are contributing to learning and other desirable outcomes (*i.e.* creativity) [5]. In *Minecraft*, an observational study examining learner understanding of urban planning, found that learners from a small town in Brazil incorporated their own interpretations of what matters in a habitat work product and included additional spaces they deemed important, such as playgrounds [1]. Assessing student *Minecraft* builds to provide in-activity learner support, however, has unique challenges. It is an ill-defined domain and there is no one right way to build a habitat. Students may incorporate structures that have personal meaning that may arise from student prior experiences [1], which can be difficult for humans, and as a result, computers to interpret. We seek to understand how AI models can be designed around digital making assessment using an exploratory approach and expand the literature regarding student support in open-ended learning activities.

3 Methods

3.1 Participants

A total of $n = 48$ middle school age students are included in this study (31% female) with an average age of 11.96 years old from camp data collected in 2022. All students participated in 1-week summer camps held in three distinct locations in the West, Midwest, and East United States. Demographic breakdown is as follows: 30% Caucasian, 23.75% African-American, 21.25% preferred not to answer, 12.5% Hispanic, 2.5% Asian, 1.25% American Indian, and 7% Other. A total of $n = 24$ bases were analyzed for the scoring scheme. To optimize our AI models using the scoring scheme, additional bases were collected in 2023. We used data from a total of $n = 131$ students that made 57 bases across 16 camps (25.69% female, 2.75% non-binary). Participants had an average age of 11.49 years old. Of the students that entered demographic information on our survey, the breakdown is as follows: 54.4% Caucasian, 20% African-American, 8% preferred not to answer, 8% Hispanic, 4% Asian, and 5.6% Other. Consent to participate was obtained from at least one parent/guardian and verbal assent was assessed at the beginning of each camp. Participant familiarity with building in Minecraft varied considerably.

3.2 Materials

Participants were all provided with a laptop, mouse, and an individual loaner (anonymous) account to play *Minecraft: Java Edition*. Participants used the same account for each session of their respective after school program or summer camp. All maps explored by participants were created by our lab and represent simulations of "What if" questions, such as "What if Earth was a moon to a larger planet?", as well as known exoplanets, planets outside of our solar system (*e.g.* Kepler 186-f). Design of worlds was done in consultation with an astrophysicist and each features extreme conditions, such as high winds, widespread volcanic activity, freezing temperatures, or low gravity, which can be seen or measured using in-game science tools. As part of the camp curriculum, participants complete self-explanation questions following their exploration of each in-game world. Each world has three total questions, scored on a scale of 0 to 3, each showing a level of astronomy explanation and comprehension of the material presented.

3.3 Procedure

During the final sessions of camps participant groups were formed based on seating arrangements, existing friendships or by researcher assignment. They were prompted with an introductory video and presentation and then challenged to design a habitat for humans to work and survive on Mars. Participants were invited to employ knowledge they gained from exploring previous hypothetical Earths and exoplanets to inform how to respond to extreme conditions on Mars. Groups had around 3 hours to collaborate and build their habitats and presented them to peers and parents on the last day, explaining the problems they addressed and how they solved them, as well as what made their habitat special.

3.4 Data Analysis

Habitat Scoring Scheme. The scoring scheme for the Mars habitats was drafted in consultation with a professor of astronomy. It was designed to account for the scientific considerations that could be represented in student bases during the Mars habitat challenge. In total, 11 categories were outlined: area where the base is built, atmosphere regulation within the base, combating different levels of gravity, communications facilities, food and water considerations, health and wellness facilities, transportation (*e.g.* rovers, rocket launchpad, etc.), power generation (*e.g.* nuclear reactors), protection from radiation, rounded structure shape, and storage for supplies. Each category was scored using a three-tier system that was later used as labels for the AI classifiers described in later sections. These tiers are classified from least score to highest score as "Basic", "Intermediate", and "Mastered", each representing a level of application and mastery that the participants show during the Mars habitat activity. The "Basic" tier awards 0 points for the category and is represented in habitats with the concept not being present or present but highly unrealistic. The "Intermediate" tier awards a number of points halfway between 0 and the maximum for each category and is represented in habitats with the concept being present within the habitat, but unfinished. The final "Mastered" tier awards the maximum number of points per category, and represents that the team integrated the concept clearly and accurately reflects what scientists would realistically use when making a real habitat on Mars.

To complete the scoring for all 24 habitats, two researchers reviewed each habitat and scored them. One researcher scored all of the habitats. The other was trained by scoring five bases individually. The remaining 19 habitats were double scored independently. Comparing scores, an average agreement of 93% emerged, with a calculated Cohen's kappa of $\kappa = 0.87$, indicating excellent agreement.

Artificial Intelligence Architecture. The habitat scores were used as labels to train AI models. The architectures for the models were chosen around the capabilities of the learning environment and the habitat scoring categories outlined above. Plug-ins on the *Minecraft* server automatically collect instances where students place or remove blocks (referred to here as block data, including the type of material, its coordinates and state) and can take screenshots of the world in-game. It is impossible, however, for a model to predict all of the categories solely from one input type (food sources cannot be interpreted from aerial images, area of the base cannot be interpreted from underground images, shape cannot be interpreted by block data, etc.). Thus, we took an exploratory approach to determine which frameworks and data sources work best for *Minecraft*.

We designed three baseline models for the three input types: aerial images, underground images, and block data. For both aerial images and underground images, we used a convolutional neural network (CNN) that consisted of 2 2-dimensional convolutional layers, batch normalization layers, dropout layers, and 2 dense layers. Images in our dataset were resized to 128x128 and each pixel value was normalized. For block data, the columns for the dataset were the types of blocks used for all of the groups and for each cell were the number of the block type used by the group normalized. We used a multi-layered perceptron (MLP) for the block data, which consisted of 3 dense layers and dropout layers.

There were a total of four multi/mixed-input classifiers that consisted of all permutations of the input types. To concatenate the models into a single classifier, we used late fusion to concatenate the output layers together to use as input to a dense connected layer to get the final classification for the category. Before concatenation, we used the same model architectures in the multi/mixed-input classifiers as the ones used in the baseline models. Each model was used to predict the score on the habitat scoring scheme for a single category.

AI Model Comparison. A total of $n = 21$ bases were used to compare different input types for the AI models, 3 were omitted from the dataset due to missing block data. To compare the seven AI frameworks described above, we used 5-fold cross-validation. To handle dataset imbalances we used class weighting and to prevent overfitting we used early stopping. We then ran an Analysis of Variance (ANOVA) to determine whether the means of the seven models for all of the categories were identical and a post hoc Tukey's Honest Significant Difference (HSD) test to determine which pairwise mean comparisons between conditions yielded significant differences.

4 Results

4.1 Habitats and Learning

To demonstrate that the scoring process was reliable for all teams, two Spearman's rank correlations were performed comparing habitat scores to team sizes and face-to-face time. There were non-significant Spearmans rank correlations between team size and habitat scores ($r[22] = -0.03$, $p = 0.89$) and face-to-face camp instruction time and habitat scores ($r[22] = 0.13$, $p = 0.54$). To assess validity, there was a significant positive correlation between group mean self-explanation score and habitat scores, $r(22) = 0.51$, $p = 0.01$.

4.2 AI Model Comparison

Model Metrics. One-way ANOVAs were conducted to compare the conditions on accuracy, precision, recall, and F1 scores. The ANOVAs for all four metrics were significant: accuracy ($F[6, 378] = 9.82$, $p < 0.01$), precision ($F[6, 378] = 6.99$, $p < 0.01$), recall ($F[6, 378] = 3.42$, $p < 0.01$), and F1 score ($F[6, 378] = 6.93$, $p < 0.01$). Table 1 shows the Tukey HSD test results of the comparisons between conditions on the metrics.

Table 1. Post hoc Tukey HSD mean differences for the conditions on the metrics (A = Aerial, U = Underground, B = Blocks). $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Comparison	Accuracy	Precision	Recall	F1
U vs. A	.04	.01	.01	.01
B vs. A	.22***	.18**	.09	.16**
A+U vs. A	.05	.02	−.01	.01
A+B vs. A	.20***	.13	.08	.13*
U+B vs. A	.19***	.15*	.09	.14*
A+U+B vs. A	.25***	.19***	.12*	.18***
B vs. U	.18**	.17**	.08	.15**
A+U vs. U	.01	.01	−.02	−.001
A+B vs. U	.16**	.12	.07	.11
U+B vs. U	.16*	.14*	.08	.12
A+U+B vs. U	.21***	.19***	.11	.17**
A+U vs. B	−.17**	−.16**	−.10	−.15**
A+B vs. B	−.02	−.05	−.01	−.03
U+B vs. B	−.03	−.03	−.004	−.02
A+U+B vs. B	.03	.02	.03	.02
A+B vs. A+U	.15*	.11	.09	.12
U+B vs. A+U	.14*	.13	.10	.12
A+U+B vs. A+U	.20***	.18**	.13*	.17**
U+B vs. A+B	−.01	.02	.01	.01
A+U+B vs. A+B	.05	.06	.04	.05
A+U+B vs. U+B	.06	.05	.03	.05

4.3 Model Optimization

A total of $n = 57$ bases were used for feature selection and finding the optimal amounts of layers and nodes for block-only models. We optimized the block-only models because they did not perform significantly worse than our best models and they could reduce deployment issues such as server lag and collecting non-representative image data using automated image capturing. For feature selection, we removed highly correlated building materials, using $r \geq .95$ as a cutoff, and kept a single column to represent the removed columns. Then, using a randomized search with 50 iterations with varying numbers of hidden layers (minimum of 1 and maximum of 4 layers) and nodes in each layer (minimum of 25 and maximum of 500 nodes), we ran a 5-fold cross validation for each habitat scoring category (results shown in Table 2).

Table 2. Results of the randomized search 5-fold cross validation for the best models regarding accuracy for each category.

Feature	Accuracy	Precision	Recall	F1
Oxygen Production/Atmosphere Regulation	.72	.57	.45	.81
Radiation Protection	.62	.57	.41	.64
Power Generation	.77	.79	.58	.77
Communication	.74	.78	.57	.67
Shape of Structure	.62	.54	.45	.72
Area Built	.76	.87	.58	.72
Transportation	.61	.57	.47	.64
Combating Lack of Gravity	.56	.55	.55	.54
Food and Water	.49	.48	.35	.60
Supplies	.67	.63	.47	.65
Health and Wellness	.62	.57	.47	.64
Average	.65	.63	.47	.68

5 Discussion

In this paper we discuss our research to develop a scoring scheme to assess student Mars habitats and a habitat classifier based on that scoring scheme. Consistent with our first two hypotheses, we conclude that the scoring method for the Mars habitats provides a fair and reliable means of assigning a numerical score to any given base built by camp participants. Because team size and face-to-face time had a non-significant impact on habitat scores, we can infer that varied contexts do not substantially impact learners' abilities to construct meaningful habitats. This was further proven by correlating the mean self-explanation score with the habitat scores. The higher a group was able to score on the questions, the better their habitat scored overall. This indicates that when participants integrate more astronomy knowledge from the exploration phase into their builds, habitats tend to be more comprehensive and accurate.

Contrary to our third hypothesis, a model built with a combination of data sources (aerial and underground images) was significantly outperformed by the model using only block data. This result is surprising, since this approach provides the same visual information that human scorers are given when assessing habitats. As noted previously, students may build representations of structures based on their prior experiences. Our findings indicate that although student builds appear different based on student background, structures that serve specific purposes use similar materials. Thus, utilizing block data models provides the necessary context to understand such builds. The dataset on which we base our study, however, is comprised mostly from Caucasian male participants, and it is unclear how generalizable our models are. To ensure our models are able to accurately assess build from diverse populations it is important that we continue to conduct member checks (such as the student habitat presentations), have scorers with backgrounds that align with our participants, and gather more data from underrepresented students in our dataset.

6 Future Work

Our habitat analysis and assessment models have been recently integrated into pedagogical agents on our *Minecraft* server. They will be used in upcoming studies to provide real-time habitat feedback with students teams during camps, which we believe will help to alleviate teacher burden and support student learning. Other future directions include providing students with a progress checklist during the activity (to identify and work on what they may have missed), designing collaborative building agents to expedite work, and integrating the system into curriculum as a stealth assessment or tool to inform student support needs.

Acknowledgments. The materials used in this study are based upon work supported by the National Science Foundation and Directorate for Education and Human Resources under Grants 1713609 and 1906873.

References

1. de Andrade, B., Poplin, A., Sousa de Sena, Í.: Minecraft as a tool for engaging children in urban planning: a case study in Tirol town, Brazil. ISPRS Int. J. Geo-Inf. **9**(3), 170 (2020)
2. Harel, I.E., Papert, S.E.: Constructionism. Ablex Publishing (1991)
3. Lane, H.C., et al.: Triggering stem interest with minecraft in a hybrid summer camp (2022)
4. Nyhout, A., Ganea, P.A.: Scientific reasoning and counterfactual reasoning in development. Adv. Child Dev. Behav. **61**, 223–253 (2021)
5. Papavlasopoulou, S., Giannakos, M.N., Jaccheri, L.: Empirical studies on the maker movement, a promising approach to learning: A literature review. Entertain. Comput. **18**, 57–78 (2017)

6. Sharma, K., Giannakos, M.: Multimodal data capabilities for learning: what can multimodal data tell us about learning? Br. J. Edu. Technol. **51**(5), 1450–1484 (2020)
7. Woodward, M.: Minecraft user statistics: How many people play minecraft in 2023? (2023). https://www.searchlogistics.com/learn/statistics/minecraft-user-statistics/