Fine Grain Synthetic Educational Data: Challenges and Limitations of Collaborative Learning Analytics

Flanagan, Brendan; Majumdar, Rwitajit; Ogata, Hiroaki

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 3, 2022, accepted February 17, 2022, date of publication March 3, 2022, date of current version March 11, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3156073

Fine Grain Synthetic Educational Data: Challenges and Limitations of Collaborative Learning Analytics

BRENDAN FLANAGAN, (Member, IEEE), RWITAJIT MAJUMDAR, (Member, IEEE), AND HIROAKI OGATA, (Senior Member, IEEE)
Academic Center for Computing and Media Studies, Kyoto University, Kyoto 606-8312, Japan

Corresponding author: Brendan Flanagan (flanagan.brendanjohn.4n@kyoto-u.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (B) under Grant 20H01722, in part by JSPS Grant-in-Aid for Scientific Research (Exploratory) under Grant 21K19824, in part by JSPS Grant-in-Aid for Scientific Research (S) under Grant 16H06304, and in part by the New Energy and Industrial Technology Development Organization (NEDO) under Grant JPNP20006 and Grant JPNP18013.

ABSTRACT While data privacy is a key aspect of Learning Analytics, it often creates difficulty when promoting research into underexplored contexts as it limits data sharing. To overcome this problem, the generation of synthetic data has been proposed and discussed within the LA community. However, there has been little work that has explored the use of synthetic data in real-world situations. This research examines the effectiveness of using synthetic data for training academic performance prediction models, and the challenges and limitations of using the proposed data sharing method. To evaluate the effectiveness of the method, we generate synthetic data from a private dataset, and distribute it to the participants of a data challenge to train prediction models. Participants submitted their models as docker containers for evaluation and ranking on holdout synthetic data. A post-hoc analysis was conducted on the top 10 participant's models by comparing the evaluation of their performance on synthetic and private validation datasets. Several models trained on synthetic data were found to perform significantly poorer when applied to the non-synthetic private dataset. The main contribution of this research is to understand the challenges and limitations of applying predictive models trained on synthetic data in real-world situations. Due to these challenges, the paper recommends model designs that can inform future successful adoption of synthetic data in real-world educational data systems.

INDEX TERMS Synthetic learner data, student modeling, data sharing, data challenge.

I. INTRODUCTION

As educational systems are collecting an increasing amount of data on the learning behavior of students, its analysis has given rise to the fields of Educational Data Mining, and more recently Learning Analytics. As the adoption of digital learning environments gathers momentum, laws such as General Data Protection Regulation (GDPR) in the European Union, and institutional based provisions such as Family Educational Rights and Privacy Act (FERPA) and the IRB Common Rule have been implemented to protect student and teacher privacy. This has given rise to the investigation of methods for de-identification [1], policies [2], frameworks [3], [4], and platforms [5], [6] to protect data ownership rights while fostering research into how educational data can be analyzed to improve the effectiveness of learning systems and teaching practices.

Ethical use and sharing of data have been key issues for learning analytics since its inception [7]. Privacy and ethical risks of data sharing have been identified as an important issue through a survey of researchers and practitioners conducted by a European learning analytics support action [8]. Furthermore, previous studies found that students in higher education are conservative when it comes to sharing data [9]. While data privacy is of upmost importance in learning analytics, the protection of personal data often creates difficulty when promoting research into underexplored contexts as it can inhibit the sharing of data with researchers outside of institutions and core project groups. It is also acknowledged within the research community that data sharing for the purposes of research replication, interoperability and fostering further research development is critical to broadening the

The associate editor coordinating the review of this manuscript and approving it for publication was John Mitchell.

acceptance and adoption of learning analytics [10], [11]. Fischer *et al.* [12], suggest that while there are inherent risks with sharing learner data due to privacy and personal information, there are also risks for not sharing learner data due to strict policies, such as: comparing and evaluating educational institution performance, and the impact that educational programs have on academic performance. This has led to discussion on how to overcome such problems and several proposals [13], including the generation of synthetic data. The later method has been applied to enable greater scope in analyzing longitudinal data that would otherwise be inaccessible due to government data sharing policies [14]. However, to date there has been little work that has explored the adoption and real-world limitations of synthetic data when sharing with the wider research field. Ferguson *et al.* [15] proposed a check list consisting of 21 challenges and ethical dimensions in learning analytics, including ''Share insights and findings across digital divides'', ''Anonymize and de-identify individuals'', and ''Provide additional safeguards for sensitive data''. While sharing insights and findings are important for progressing the field of learning analytics, the sharing of important datasets across digital divides can support greater inclusiveness within the community by providing access to data that would otherwise be inaccessible to a wide range of researchers. The successful use of synthetic data is a promising solution to the key problem of sharing anonymized sensitive data to the broader learning analytics community.

Student performance prediction research aims to identify learners who could benefit from early intervention to mitigate low academic performance or course drop-out [16]. The analysis of these research is drawn from pre-course state such as: socio-economic, psychological traits, questionnaire [17], and past performance [18], [19], behavioral data from interactions with learning systems [20]–[22], and external systems such as social media [23], and also multisource data [24]. Investigation into student performance prediction models has mainly focused on higher education with few works targeting K-12 [25]. In addition, the type of data analyzed for prediction has tended to target socio-economic or pre-course performance, with less focus on fine grain behavioral data which is presented in the current study.

This paper examines the effectiveness and challenges of creating and distributing a synthetic dataset that has been generated from a private dataset that would otherwise not be distributable due to data privacy restriction policies. To investigate this method, a dataset consisting of two types of data: reading behavior data that was collected from a digital reading system [6] and the final academic performance scores, was used to train a generative model. This model was then used to generate a synthetic data for distribution to third parties. The authors conducted a data challenge to recruit various third parties from research and the private sector to train prediction models on synthetic data. This ensured that there was competition between the data challenge participants to train high performing prediction models. Participants were given the task of predicting the academic performance of learners based on the analysis of their reading behavior, and were encouraged to submit models that had been constructed by analyzing the distributed synthetic dataset. For the purposes of the data challenge, the models were evaluated initially on a holdout synthetic dataset generated using the same technique. Post-hoc analysis of the submitted models was performed using the original private dataset to investigate the effectiveness of implementing the sharing of synthetic data to third parties to construct prediction models and effectiveness in deploying models in a real-world scenario.

The novel contributions of this paper are summarized as follows:

- We propose a method of generating educational synthetic data for a digital learning material reading system and students final scores.
- A comparison of academic achievement prediction models trained on real and synthetic data is conducted to verify the effectiveness of the method.
- The synthetic data was provided to third parties as a part of a competitive data challenge to construct early warning prediction models. This shows promising results from data sharing.
- Discussion of challenges and limitations that led to poor model transfer from a practical implementation of synthetic data use.
- Propose recommendations on model design that can inform future successful adoption of synthetic data in real-world educational data systems.

II. RELATED WORK

The analysis and use of artificial data in education consists of two main branches: simulated learners and synthetic data. Simulated learners are often used when it is not possible for actual learners to use the system, and a model of anticipated learning behavior is constructed as a proxy, such as: when a learning system is in the conceptual, development, or testing stage where it is still unknown how interaction with the system will unfold, or as agents in the education of actual learners or teachers. Synthetic data on the other hand is generated from a model trained on actual data collected from real user interactions with the learning system. The latter can be employed to overcome a range of different issues relating to data use and privacy concerns.

A. SIMULATED LEARNERS

There is a long history of using simulated learners in educational technology for a range of different purposes, with the seminal research by VanLehn *et al.* [26] proposing simulated learners as a tutor training system for teachers, learning partners for students, and a testing ground for instructional designers to perform formative evaluation of existing learning systems [27]. While simulated learners draw some parallels with pedagogical agents in that they can populate and play a role in learning systems, but they can also be used to examine problems with the design and use that could potentially be too costly or not possible to implement with human subjects [28]. The simulation of learner data has also been put forward as a possible solution for investigating problems at scales which might otherwise be impossible to examine due to lack of data that can be ethically collected [13]. In student modeling, the role that small differences can play in prediction performance have been investigated by modifying characteristics of simulated learners [29]. Simulated students have also proved useful for examining the inner workings of knowledge tracing models and support the testing of hypotheses on appropriate evaluation metrics which could translate directly to better correlation with improvements of knowledge estimation [30]. While the role of simulated learners will continue to play an important role in artificial intelligence in education and learning analytics, there are some limitations that should be considered. Simulations are based on theoretical assumptions use to inform the design of models that generate simulated interactions, and real students could behave differently due to real-world factors such as affect and fatigue [31]. This could potentially limit the utility of such methods, especially for exploratory investigation, and modeling of student behaviors to predict learning outcomes.

B. SYNTHETIC DATA

The concept of synthetic data has played an important role in enabling access to data, with early incarnations of methods used to anonymize confidential tabular microdata as proposed by Rubin [32]. As use of internet-based services increased, interest in temporal and sequential synthetic data generation gained greater attention with the use of sequence networks and other methods to model and generate clickstream data for the training of personalized recommender systems from private data where privacy policies restricted publication and sharing [33]. More recently with the success of deep neural network-based models there has been renewed interested in the use of synthetic data in many fields ranging from automated driving algorithms [34] to medical imaging [35]. Greater use of synthetic data has also prompted the discussion of methods for evaluating the appropriateness of the generated synthetic data and effectiveness for use. El Emam [36] proposed several methods for evaluating the utility of synthetic data, including the structural similarity to the original data, general utility metrics, bias and stability assessment. Based on these guidelines, we ensure that the generated synthetic data in this paper has the same structural composition, and conduct a preliminary evaluation to measure the utility of the data for the proposed purpose.

The rising use of synthetic data has also prompted research into general methods for generating synthetic data. Domain agnostic methods learn the relation of different features from the private dataset, such as DataSynthesizer [37] which constructs Bayesian networks to model correlated features from tabular datasets. These methods are often applicable to tabular data, such as: demographics. Recent research has also examined generic methods for spaciotemporal data, such as SynSys [38] which constructs a generative model based on hidden Markov models and regression models based on a private dataset from sensors used for at home healthcare intervention. While these techniques may be applicable to a wide range of problems in which synthetic data can be applied, El Emam [36] suggests that the structure of data in complex systems and its analysis is often domain specific in nature, therefore making it difficult to use generic methods.

The use of synthetic data in learning analytics over the last half decade has been growing, with a two-day hackathon at LAK16 focusing on synthetic data for applications ranging from large scale performance testing to developing central data governance practices [39]. The concept of peer reviewed synthetic data generators was proposed, and these could be used to generate standardized datasets by applying repeatable recipes that could be used for the distribution of data between education institutes, encouraging the reproduction of results. It was also proposed that this could be developed at the infrastructure level to address ethical and privacy risks throughout the service cycle. Dorodchi *et al.* [40] proposed using generic off the shelf systems to generate synthetic data from demographic educational data. A Random Forest classifier was trained on both datasets, and resulted in a drop of 6.6% and 0.07 for accuracy and F1 respectively. However, this method is limited to categorical or continuous features and could not be applied to time-series data that is often the target of recent studies. Peña-Ayala [41] suggests that while there are various works in learning analytics into the prediction of student dropout and at-risk prediction, there is also an equal balance of research that looks into issues to effectively support these tasks, such as sourcing data from multimodal systems, training, testing, synthetic data generation, and deployment to applications. This shows the importance of techniques that are required to support key tasks within the field.

C. COLLABORATIVE LEARNING ANALYTICS

Collaborative learning analytics is defined as a partnership between educational researchers, data scientists, and practitioners, as well as those from various fields that share the same goal of improving learning based on data [42]. Experts from these fields need to work together to examine the full breadth of learning analytics and educational data mining [43], however few research institutes have experts from each field working closely together and sharing data.

Previous research into supporting collaborative learning analytics has examine methods of study replication and analysis of private data. Berg *et al.* [39] discussed data sharing in two case studies of an institution wide project, UvAINFORM, and the Jisc learning analytics architecture which would share data at the regional or national level between participating organizations. While these case studies aim to enable data sharing, it is limited to participants within a closed group. Gardner *et al.* [44] proposed the MORF framework to abstract data from the development of predictive modeling, allowing researchers to submit their work to a server and evaluating the effectiveness on raw MOOCs data that would otherwise not be available for analysis due to privacy restrictions.

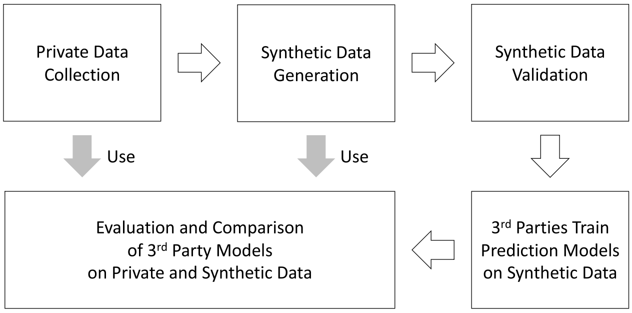

Private Data Collection
Synthetic Data Generation
Synthetic Data Validation
Use
Use
Evaluation and Comparison of 3rd Party Models on Private and Synthetic Data
3rd Parties Train Prediction Models on Synthetic Data

FIGURE 1. Research procedure diagram.

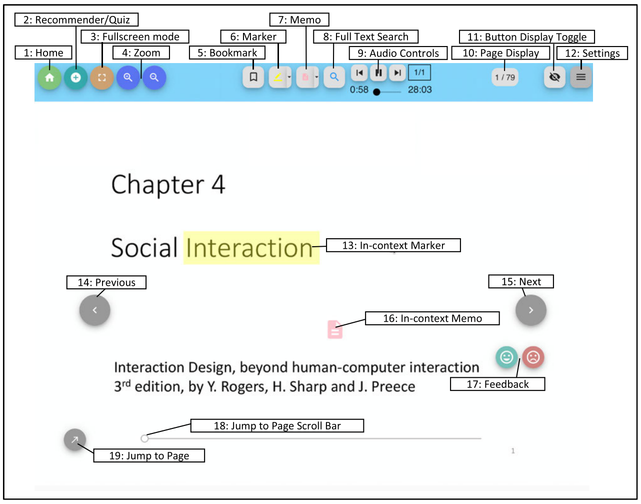
The main aim of the framework is to support replication studies and transparency for student modeling research. Submissions are made as a batch job in the form of a docker container image, python controller script, and a job metadata file, with the results and container image being automatically published and metrics are sent via email. This enables the analysis and replication of experiments on sensitive data, however there are several limitations that still exist, such as: difficulty of use due to batch submission of experiments as docker containers, and security issues of handling non-anonymized data. Also, researchers may have to change their workflow to accommodate the requirements of the framework, making it inefficient to conduct analysis. As noted by Fischer *et al.* [12], because of the challenges involved in pursuing a collaborative learning analytics approach, there are few projects where data is shared by adopting the open science values in an interdisciplinary team. Addressing these challenges, such as balancing privacy and data sharing is important because of the potential that can be achieved through big data in education.

III. METHODS

In this paper, we propose a method of creating a model to generate synthetic data trained on the characteristics of a private dataset. As the main objective for using this method is to broaden the scope of researchers that can have access to and analyze the data, without impeaching on the personal information rights of learners and teachers. An overview of the research procedure, including the methods and experiments flow is shown in Figure 1. First, we collect the reading behavior log data and assessment data from the learning systems as a private dataset that is not shared with third parties. Second, a generative model is trained on the private dataset, and then the model is used to generate the synthetic dataset. The appropriateness of the synthetic dataset for predicting academic performance is then validated. The synthetic dataset is then distributed to third parties who train prediction models on the synthetic data. Finally, a comparison of the evaluation of the models is conducted using the private and synthetic data to evaluate the effectiveness of using shared synthetic data to train prediction models.

A. LEARNING SYSTEM FOR DATA COLLECTION

Digital learning material reading systems are a core part of modern formal education. Recently, in many countries around the world there has been a push toward digitizing learning materials, and in particular existing paper textbooks and exercise books. In Japan this has been happening at the government level, with a plan to deploy nationwide digital learning material reading systems to compulsory education and abolishing paper-based textbooks [45]. For this reason, we decided to focus on reading behavior data as such systems are playing an increasingly important role in education. In addition to serving as a learning material distribution platform, it is also an important source of data for learning analytics into the reading habits of students. The action events of the readers are recorded, such as: turning to the next or previous page, jumping to different pages, memos, comments, bookmarks, and markers indicating parts of the learning materials that are hard to understand or are of importance. The reading behavior of students has previously been used to visualize class preparation [45] and review patterns [46]. A digital learning material reading system can be used to not only log the actions of students reading reference materials, but also to distribute lecture slides.



Chapter 4

Social Interaction

Interaction Design, beyond human-computer interaction
3rd edition, by Y. Rogers, H. Sharp and J. Preece

FIGURE 2. The user interface of the BookRoll digital learning material reader.

In the present work, the non-proprietary BookRoll digital learning material reading system [6] was used to serve lecture materials and capture learners reading behavior for analysis. As shown in Figure 2, the user interface supports a variety of functions, including navigation and annotation, audio narration playback, and features for measuring the learners internal state through feedback.

Currently, learning material content can be uploaded to BookRoll in PDF format, and it supports a wide range of devices, including: notebook computers, tablets, and smartphones, as it can be accessed through a standard web browser. Reading behavior while using the BookRoll system is sent using the xAPI standard in the form of a pseudonymized learning event logging and collected in an LRS. Learners can access BookRoll from the course site on the educational institutions LMS via LTI (Learning Tools Interoperability).

TABLE 1. A sample of events recorded from user interaction with BookRoll.

Contents id	Memo text	Operation date	Operation name	Page no	User id
EBOOK_341		2018/01/22 18:10	REGIST CONTENTS	0	t1
EBOOK_341		2018/01/23 9:16	OPEN	1	s1
EBOOK_341		2018/01/23 9:20	NEXT	2	s1
EBOOK_341		2018/01/23 9:21	OPEN	1	s2
EBOOK_341	Sample memo	2018/01/23 9:22	ADD MEMO	2	s1

TABLE 2. Operation names and descriptions for learning behavior interactions captured with BookRoll.

Operation Name	Description
OPEN	opened the book
CLOSE	closed the book
NEXT	went to the next page
PREV	went to the previous page
PAGE_JUMP	jumped to a particular page
ADD BOOKMARK	added a bookmark to current page
ADD MARKER	added a marker to current page
ADD MEMO	added a memo to current page
CHANGE MEMO	edited an existing memo
DELETE BOOKMARK	deleted a bookmark on current page
DELETE MARKER	deleted a marker on current page
DELETE_MEMO	deleted a memo on current page

The data collected by BookRoll is a simple event log of the users' interactions with the system, and Table 1 presents a sample of learner behavior logs that have been extracted from an LRS into tabular format. This data can be processed further to extract information about how long learners are spending on reading and other tasks while using the BookRoll system [47].

In the logs there are many types of operations which represent different interactions with the BookRoll system, for example, "OPEN" means that the student opened an e-book and "NEXT" means that he or she clicked the next button to move to the subsequent page. An overview of the types of operations and description of the interaction that is represented is shown in Table 2. A system was proposed by Flanagan & Ogata [6] that defined a framework for a learning analytics platform that can collect learner behavior data similar to that which is analyzed in the present research.

The learning system on which the proposed method is built on focuses mainly on learning, revising, and assessment which is provided in context within the BookRoll digital learning material reading system.

B. CHARACTERISTICS OF THE PRIVATE DATASET

The private dataset was collected using the LEAF platform [6] shown in Figure 3 at a secondary school in a mathematics class over the course of 4 months in a semester in early 2020. All students were provided a notebook computer to use the system and were able to read the materials in class and also outside the classroom. The learning materials that were read using the BookRoll system ranged from textbooks to teacher created handouts, contents and worksheets. In addition to the reading behavior data, academic achievement data of the students in the class during the collection period was also provided by the school, consisting of periodic tests and final exam scores. These scores were aggregated using the method use to calculated a single academic achievement score by the school that was normalized to a scale from 0 to 100. The reading behavior logs were filtered to remove data of students that did not consent to data collection, or had missing academic achievement data due to absences. A total of 120 students' data was collected, consisting of 65,387 reading behavior logs.

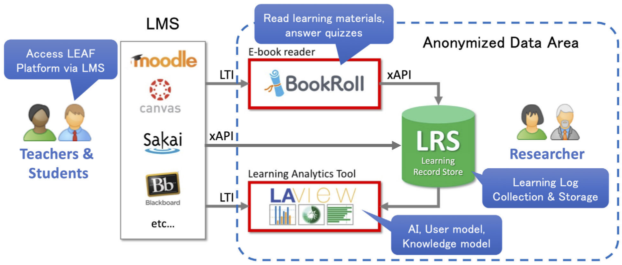


FIGURE 3. Overview of the LEAF Platform used for data collection.

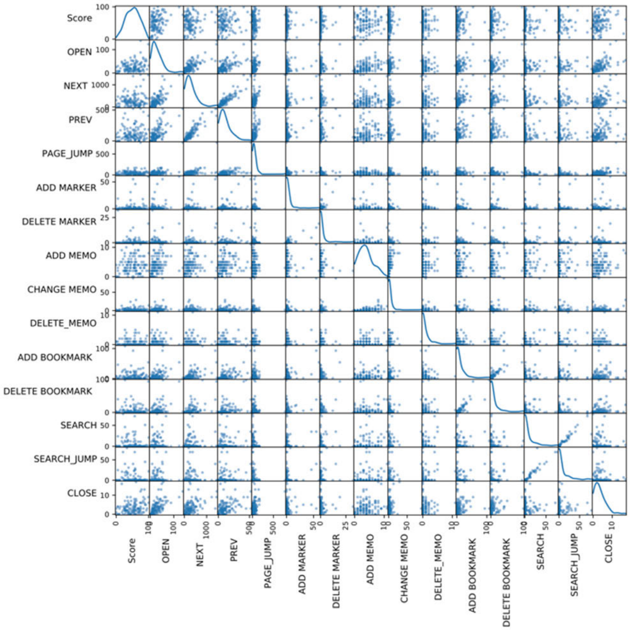


FIGURE 4. Distribution and pair plots of academic performance (score) and frequency of reading behavior operations.

In Figure 4 the distributions and pair plots of reading behavior operations and the academic achievement score from the original data is shown. It can be seen that the scores have a relatively normal distribution when compared to operation frequency, which have mainly long tail distributions.

TABLE 3. Relationship of academic performance (score) and frequency of reading behavior operations.

	Score	OPEN	NEXT	PREV	PAGE_ JUMP	ADD MARKER	DELETE MARKER	ADD MEMO	CHANGE MEMO	DELETE_ MEMO	ADD BOOKMARK	DELETE BOOKMARK	SEARCH	SEARCH_ JUMP
OPEN	0.3***													
NEXT	0.17	0.75***												
PREV	0.13	0.67***	0.91***											
PAGE_JUMP	0.11	0.54***	0.54***	0.52***										
ADD MARKER	0.12	0.28**	0.43***	0.43***	0.02									
DELETE MARKER	-0.04	0.13	0.26**	0.22*	0	0.7***								
ADD MEMO	0.12	0.31***	0.18	0.23*	-0.03	0.1	0.02							
CHANGE MEMO	0.05	0.19*	0.33***	0.33***	-0.03	0.43***	0.27**	0.5***						
DELETE_ MEMO	0.12	0.04	0.05	0.08	-0.05	0.08	0.01	0.17	0.13					
ADD BOOKMARK	-0.01	0.42***	0.59***	0.58***	0	0.43***	0.23*	0.2*	0.47***	0.09				
DELETE BOOKMARK	0.03	0.44***	0.5***	0.44***	-0.01	0.43***	0.27**	0.09	0.19*	0.11	0.87***			
SEARCH	-0.09	0.1	0.06	0.05	-0.17	0.13	0.05	0.17	0.23*	0.27**	0.1	0.05		
SEARCH_ JUMP	-0.11	0.08	0.06	0.05	-0.17	0.13	0.09	0.16	0.23**	0.23*	0.12	0.07	0.94***	
CLOSE	0.24**	0.62***	0.43***	0.39***	0.37***	0.12	0.06	0.23*	0.13	0.02	0.16	0.19*	-0.03	-0.05

* p < .05, ** p < .01, *** p < .001

It should be noted that BookRoll will always record an 'OPEN' operation when an ebook is opened from the list of contents, however if a student completes a reading session by terminating the web browser it is not possible to record a 'CLOSE' operation and therefore there is a large difference in the frequency of these two operations.

The correlation of academic performance and frequency of reading behavior operations is shown in Table 3. It should be pointed out that the only operations that have significant weak correlation with academic performance are the 'OPEN' and 'CLOSE' operations which indicates that students who have frequent reading sessions tend to have higher academic achievement. Some of the significant correlation between operations could be explained by user interface design choices, such as 'SEARCH' which represents text searching within a material then 'SEARCH JUMP' which indicates that a user selected a search result.

IV. SYNTHETIC DATA GENERATION

There are several aspects that need to be taken into account when training a model to generate synthetic reading behavior data that might otherwise not be applicable in other behavioral or temporal [48] data domains. One such aspect is the inherent limitations of the material that is being modeled: a digital learning material that is created in a similar manner to a traditional book or pdf. For data generation, a generative model was first trained on the original data. While recent methods of synthetic data have proposed using deep neural network-based methods such as generative adversarial networks in medical imaging classification [35], these methods often require a large amount of real data samples to effectively generate synthetic data. Due to the amount of student data available in the private dataset, we decided to adapt a first-order Markov model from the Pomegranate python package [49] to the task of generating reading behavior data as similar methods have been successfully applied to synthetic data generation in previous research [38].

Actions by the learner were represented as states in the Markov model and transition probabilities between states were learned from the original data. Reading sessions in the original data were extracted by detecting the 'OPEN' and 'CLOSE' events, periods between events that are longer than 20 minutes, or a change in the learning material that is being read. The beginning and finish of sessions were represented by the 'Start' and 'End' states in the Markov model respectively. For each student in the synthetic data simple attributes such as the timestamp of the last event generated are used to inform the timing of consecutive events. Timestamps for each reading behavior event are generated using Monte Carlo sampling to ensure that the distribution of the generated values resembles that of the original data. For example, when a reading session begins, the initial timestamp

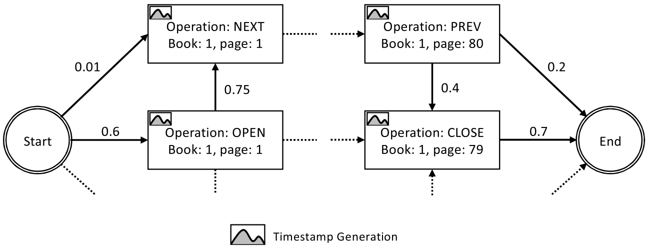


FIGURE 5. **Overview of the synthetic data generation model.**

is selected based on Monte Carlo sampling of the time of day, day of week from the last event timestamp for the student. Subsequent events timestamps are generated depending on the state of the Markov model and the distribution of the associated the timestamps from the original data. Figure 5 shows an overview of how the Markov model and timestamp generation was constructed, with each state generating a time stamp based on the previous state before the transition. As the target of the data generation is to predict the academic achievement based on reading behavior, the transition probabilities of the model were also determined according to the score that was generated for the particular student, thus capturing the differences in reading behavior of students at different levels of academic achievement.

A. FITNESS OF GENERATED SYNTHETIC DATA FOR MODEL TRAINING

To verify that the generated synthetic data is a viable alternative to the private dataset for training models, we evaluated the difference in performance of a model trained on the original data and compared it to the performance of a model trained on the generated synthetic data. In both cases the model performance was assessed by predicting the academic achievement from student data in the holdout dataset from the private data.

We based the model on previous research that predicted academic achievement from reading behavior data in the higher education context. Akçapınar, *et al.* [50] evaluated the performance of 13 prediction algorithms on aggregate features from log data and found that Random Forest had high accuracy in predicting academic achievement, and this has also been confirmed in similar studies that analyzed learning interaction data [25], [22], [19]. Based on this, we decided to train Random Forest regression models to compare the performance of using synthetic and real training data.

The features shown in Table 4 were generated from the raw log data, and the aggregate counts were normalized by percentile rank *PR* for each student as shown in the equation below, where f_b is the number of students with values less than the single student's value of the percentile rank, f_w is the number of students with values the same value as the value of the single student's percentile rank, and N is the total number of values.

$$PR = \frac{f_b + {}^{1}\!/_{2} f_w}{N} \tag{1}$$

TABLE 4. **Description of aggregate features [41].**

Features	Description
totalevent	Total number of events
content	Number of different contents studied by the student
session	Number of reading sessions by the student
time	Total time spend on eBook system in minutes
week	Number of different weeks that student use the system
day	Number of different days that student use the system
completionrate	Average completion rate of all books
longevent	Number of events longer than 3 s
shortevent	Number of events less than or equal to 3 s
next	Number of Next events
previous	Number of Previous events
jump	Number of Jump events
markerimportant	Number of important markers added by the student
yellowdifficult	Number of difficult markers added by the student
memo	Number of memos added by the student
bookmark	Number of bookmarks added by the student
score	Academic performance of students at the end of the semester

TABLE 5. **A preliminary comparison of the performance of Random Forest regressors trained on original and synthetic data.**

Training Data	Mean	SD	SE	t	df	p
Private	23.10	2.17	0.48	-1.08	19	0.29
Synthetic	23.58	1.21	0.29			

To evaluate the performance of the models RMSE of the predictions was averaged using 5-fold stratified cross validation over 20 randomized trials as proposed by Japkowicz & Shah [51]. Further, we conducted a t-test on the results of the randomized trials to test the significance of the predictions from the models trained on private and synthetic data as shown in Table 5. The private data trained model ($M = 23.10$, $SD = 2.17$) compared to the synthetic data trained model ($M = 23.58$, $SD = 1.21$) indicate that there was no significant difference of predictive performance measured by RMSE when evaluated over 20 randomized trials, $t(19) = -1.08$, $p = 0.29$. The model trained on the private data has a marginally but not significantly better performance than that of the synthetic data model, and therefore confirms that the synthetic data can be a viable alternative for training a Random Forest model to predict academic achievement on private data.

B. SYNTHETIC DATA DISTRIBUTION

A synthetic dataset consisting of 12.6 million generated reading behavior interaction logs from 10,000 unique students was distributed to 60 registered participant teams as part of a data challenge. Figure 6 is an overview of the types of data that were distributed to the teams, including both the reading activity log data, final exam score data and small subsets to be used as test data. A holdout dataset was not distributed to participants, and was used to evaluated the performance of submitted models. Participating teams ranged from higher education research institutes to private sector ICT and EdTech companies involved in providing proprietary digital learning

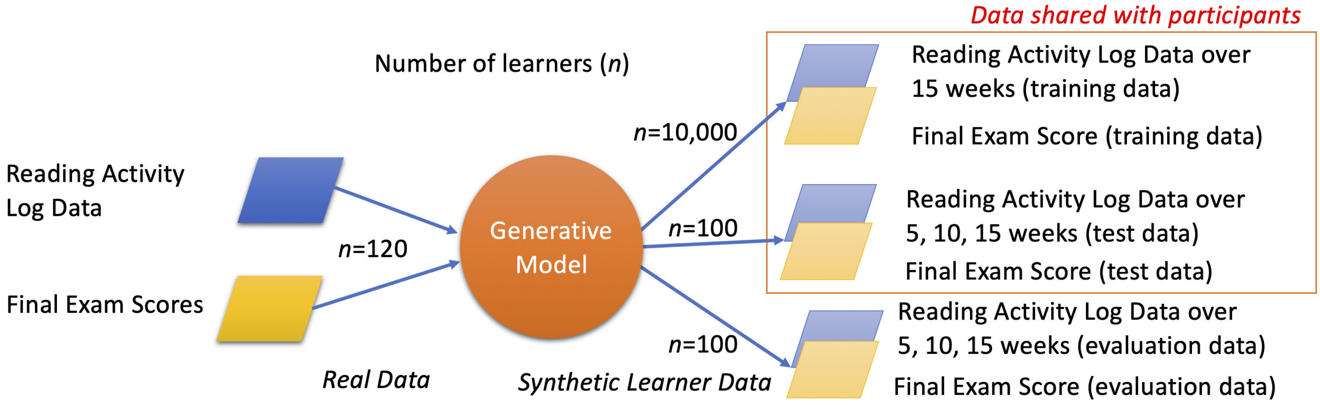

Data shared with participants
Number of learners (n)
Reading Activity Log Data
Final Exam Scores
n=120
Generative Model
n=10,000
n=100
n=100
Reading Activity Log Data over 15 weeks (training data)
Final Exam Score (training data)
Reading Activity Log Data over 5, 10, 15 weeks (test data)
Final Exam Score (test data)
Reading Activity Log Data over 5, 10, 15 weeks (evaluation data)
Final Exam Score (evaluation data)
Real Data
Synthetic Learner Data

FIGURE 6. Overview of data generation and types of data distributed.

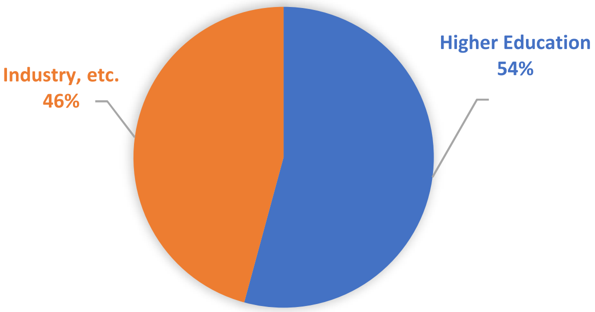


FIGURE 7. Distribution of participant association (n = 60).

platforms, and the distribution of the participants association is shown in Figure 7.

Over the span of the data challenge, teams were encouraged to submit their models as docker containers for evaluation on a synthetic holdout dataset that was not released to participants, and evaluation submissions were limited to once every 24 hours to limited the ability to tune the results for the specific evaluation. The directory structure was based defined as shown in Figure 8 to ensure compatibility when testing and replicating the evaluation of the submission on held-out data at two difference higher education institutions. The design of the container layout was based on that proposed for the MORF framework [44], and modified it to suite the data structure and data challenge requirements, such as: automated testing and evaluation at scale. The container structure allows for easy substitution of data by utilizing dockers files system mount feature to inject external files without having to modify the container directly. The app directory contains two shell files *evaluate.sh* and *train.sh* that can be run to perform a model evaluation or training respectively. Model files from training can be saved in the *model* directory for evaluation at a later time. The *data* directory contains three different types of data: *evaluate* data which is analyzed to predict student performance, *result* which contains the model prediction output, and *source* which holds synthetic data that is analyzed to train a prediction model. The schema for the *EventStream.csv* and *QuizScore.csv* files are shown in Table 6 and 7. It should be noted that the *QuizScore.csv* file in the *evaluate* directory does not contain any values for the score column. The *out.csv* file which contains the predictions of student performance should be output in the same order as the *userid* in the evaluation *QuizScore.csv* file.

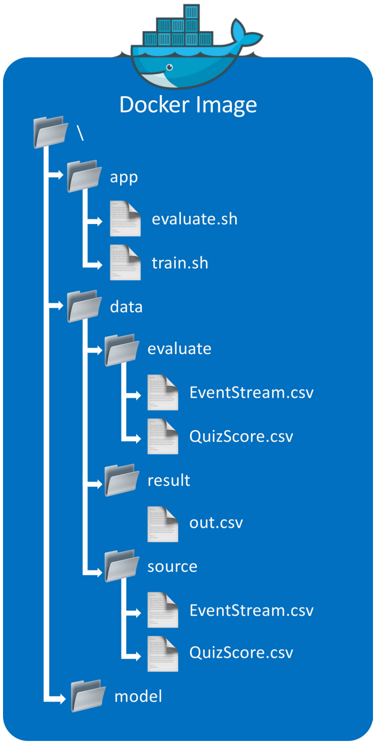


FIGURE 8. Standardized file structure of submission docker image.

V. EXPERIMENT

One of the primary goals of predicting academic performance based on learner behavior logs is to identify underperforming students that require intervention support [16], [42]. It is

TABLE 6. The schema of the EventStream.csv file.

Column	Description
userid	Anonymized student userid, eg: 335b81af-7e60-4a04-8712-ed734eccbf62
contentsid	The id of the e-book that is being read.
operationname	The action that was performed in BookRoll. (see list of operation names in Table 2)
pageno	The current page number where the action was performed.
marker	The reason for the marker added to a page. eg: important, difficult (contents are not understood)
memo_length	The length of the memo that was written on the page.
devicecode	Type of device used to view BookRoll, eg: mobile, pc.
eventtime	The timestamp of when the event occurred.

TABLE 7. The schema of the QuizScore.csv file.

Column	Description
userid	Anonymized student userid, eg: 335b81af-7e60-4a04-8712-ed734eccbf62
score	The final total score out of 100 that the student received for the course.

beneficial for an intervention to be carried out as soon as possible, and therefore models should be evaluated at different points in time during the course, and typically this is defined as a fixed period in relation to the duration of the course. Previous research on predicted academic performance from reading behavior data has found that model accuracy increases from week 3-5 onwards in 15-week higher education courses [50], [52]. Based on this it was decided that the model evaluation should be carried out at 5-week intervals from the start of the course to measure the effectiveness of models to predict academic performance. The model evaluation was performed by RMSE of the predicted score using data up to the 5th, 10th and 15th week of the data collection period. To evaluate the overall performance of models, the mean of the RMSE for all evaluation periods was calculated. The mean RMSE results of the evaluation were provided on a public leaderboard in the data challenge to encourage competition between the teams. The final data challenge evaluation results were confirmed at two Japanese national universities.

To measure the effectiveness of training models on synthetic data for real-world academic performance prediction, we designed a post-hoc test using the top 10 performing models from the data challenge as they had similar or better performance than the baseline Random Forest model used to assess the fitness of synthetic data. The pretrained models that were submitted by teams in the data challenge were evaluated both on a synthetic and original private dataset to measure the difference in performance. A large difference in performance between the synthetic and original private dataset indicates that using synthetic data to train models was not effective. Conversely, if there was little difference in prediction performance then training a model on synthetic data could be seen as an effective method of sharing sensitive private datasets.

TABLE 8. The performance of each model by RMSE for evaluation on the synthetic holdout dataset and the original private dataset, ordered by average performance on the synthetic dataset.

Team	Synthetic Dataset				Private Dataset			
	W5	W10	W15	Average	W5	W10	W15	Average
1	**10.25**	**11.76**	**10.65**	**10.89**	23.75	**23.95**	23.82	23.85
2	11.80	18.97	13.35	14.71	27.62	28.14	27.46	27.74
3	12.47	18.95	14.73	15.38	27.63	26.12	27.75	27.17
4	14.23	18.19	15.28	15.90	24.13	24.88	25.40	24.80
5	14.07	19.97	14.82	16.29	25.13	24.89	25.29	25.10
6	16.55	19.16	17.81	17.84	24.04	25.01	25.84	24.97
7	10.97	27.07	15.69	17.91	26.69	25.52	26.19	26.13
8	19.48	23.61	22.63	21.91	**22.36**	24.14	**22.20**	**22.90**
9	19.50	26.04	21.99	22.51	25.17	25.64	25.27	25.36
10	25.24	25.24	25.24	25.24	23.96	23.96	23.96	23.96

The performance of the models on synthetic and original private datasets is carried out for individual predictions at 5, 10, and 15 weeks of data, and the overall average of the performance for all periods.

VI. RESULTS

The top 10 performing models submitted by participating teams were selected from the final evaluation of the data challenge, with most models performing better when evaluated on synthetic data than the baseline Random Forest model trained on aggregate features as mentioned in the methods section. The results of the comparison between models from the top 10 teams using the synthetic data and actual data are shown in Table 8, with the left half representing the final data challenge results that were evaluated on a synthetic holdout dataset, and the right half the performance of the same models when predicting academic achievement of students in the original private dataset.

Team 1 achieved an average of 10.89 on the synthetic holdout dataset, with performance at the 5th and 15th week being better than at the 10th week. It should be noted that models from most other teams also have similar fluctuations in the three prediction evaluations. However, when the same model that was trained on the synthetic dataset is use to predict the academic achievement of students in the private dataset the evaluation is markedly worse with a mean RMSE of 23.85 which is comparable to the baseline Random Forest model trained on aggregate features. The top 3 scoring teams were asked to explain the methods they used to take the prediction problem at the data challenge results presentation. Team 1 indicated that they employed aggressive feature selection on sequence features based on n-grams and data sampling techniques to achieve their final model. The remaining 2 teams fitted neural network models for each of the three time periods and selected the model to be applied based on analysis of period to which the input data belonged. The disparity between the results is also present in the evaluation of other models as well.

Team 8 has the most consistent results with the difference between the average performance on the synthetic and private dataset being less than 1 RMSE. It should also be noted that this model also had the best performance of all of the models submitted by teams on the private dataset. Compared to other teams, the model from team 8 is relatively simple, consisting of a scikit learn linear regression model trained on aggregate feature frequency that is similar to the Random Forest model used to verify the usefulness of the synthetic data. Minimal optimization of the model is also performed, and therefore increases its generalizability to datasets other than those on which the model was trained.

VII. DISCUSSION AND LIMITATIONS

Over the last half decade within the learning analytics community there has been discussion on how research should not promote the one size fits all approach to predicting academic success, and instead opting for models that are tailored for specific courses and learning design [53]. The results presented in the present paper also highlight the need for a balance between prediction performance and generalization of models when using synthetic data. This is particularly apparent from the disparity in results of participants models when tested on the synthetic and private datasets. A range of techniques were employed by teams to enhance the prediction performance of models, and this led to unintended consequences such as selecting models that were inappropriate for the input data from a multi-model method. Also, the selection of features that are optimal for the data in which that are fit for, but inflexible when applied to unseen datasets. In particular, the use of features that are specifically coded for learning materials or tasks that might be transient in nature, such as teacher made handouts that could be specific to a learning design. This result suggests that it could be more applicable to identify reading behavior associated with particular learning materials by their semantic role in the learning design, for example the knowledge or topic which the learning material covers. This could allow for features that have semantic meaning and could be applicable to a wider range of course and learning designs. To assist in the generation of such features, data challenges or published datasets could provide a list of knowledge, topics and tasks which are associated with learning materials that were read, thus potentially providing greater flexibility of analysis and further insights that would not otherwise be possible. These suggestions are also similar to that of Pelánek, Rihák, & Papoušek [54], who highlighted the impact that data collection and publication, implementation, and evaluation can have on student models.

The results of this research have important implications for collaborative learning analytics researchers when designing data sharing methods using synthetic data techniques. While this method offers a way to share datasets within the research community, appropriate precautions should be taken to ensure that artifacts from analyzing synthetic data, such as models, are validated against private data to verify valid predictions. It may also be necessary to recalibrate models by retraining on private data before it is put to use in real world learning systems. The implementation of data challenges to promote development within the research community by using synthetic data should also consider periodic validation by both synthetic and private data. This will help to detect and avoid possible problems in inflexible model designs as seen in this research. As more studies into the use of synthetic data are conducted, a set of guidelines for model designs could be drawn from these experiences to help inform future research into educational predictive modeling.

It is important to note that the present study has several limitations. Firstly, the study was conducted with a homogeneous dataset from several classes over a period of 4 months in a mathematics course, and therefore the analysis could be influenced by the learning design or domain specific characteristic of the course. A more diverse dataset from several courses might mitigate such issues, however it could also introduce additional challenges in reliably generating synthetic dataset from a heterogeneous private dataset. Courses from different domains and learning design could vary widely in the number of materials that are provided on the reading system, such as: a teacher using the system to distribute handouts to the class each lesson, or providing a large number of reading materials for exercises such as extensive reading [55].

Secondly, due to the nature of many of the model submissions, unfortunately it was not possible to retrained all of the models using the private dataset. This was due to model programs containing functions that implemented data depended processing that caused errors or abnormalities during the process. Problems encountered ranged from hard coded reading material IDs, predetermined sets of features, sampling, to pretrained models that could not be retrained on unseen datasets. Retraining models on the private dataset could possibly have overcome the reduction in prediction performance, however the data challenge was conducted only using synthetic data. Retraining on the private dataset could also inhibit data specific tuning during the process and highlights an important problem for the field: the transition from data challenge models where performance is optimized, to real world use where generalization is important to ensure usable predictions from models. Therefore, it would be advisable to design data challenges or processes where third parties analyze synthetic data in a way to mitigate possible problems in real world application, such as: the requirement for models to be retrainable on unseen data without manual tuning or the use of heuristic methods that could be rendered invalid under unexpected conditions. In the design of a data challenge, this could involve check point evaluations on a private dataset in addition to frequent evaluations that were provided to participants in this study. This could help to uncover problems with the possible real-world adoption and generalization of models designed on synthetic data while protecting the details of the private dataset from being revealed.

VIII. CONCLUSION

This study investigated reducing limitations that are placed on private datasets by implementing real-world synthetic data sharing for the learning analytics task of predicting student academic performance. A method of generating educational synthetic data for a digital learning material reading system was proposed. A preliminary evaluation using a Random Forest model was conducted to verify the suitability of the synthetic data for training prediction models. It was found to not be significantly different from the performance of a model trained on the original private dataset.

A data challenge was conducted and third party participating teams were encouraged to submit docker containers containing their prediction model for evaluation. A standard structure of the container was implemented for reproducibility and to enable scaled-up automated evaluation. The data challenge showed promising results when evaluation was conducted on the holdout synthetic data.

However, many of the top performing models from the data challenge did not perform so well when predicting academic performance on data from the original private dataset. This highlights a possible limitation of using synthetic data sharing with third parties for the purposes of developing and constructing academic performance prediction models. Some of these limitations could be mitigated to some extent through the implementation of periodic checkpoints to verify how models generalize and transfer to original private datasets. Also, some considerations for model design, such as: avoiding the selection of a static set of limited features that inhibit generalization of the model.

The study presented in this paper focused on digital material reading system behavioral data, and there is much scope in future work for implementation of similar techniques and recommendations from findings in this paper to be applied to other parts of learning systems. The suggested use of methods for synthetic data sharing in collaborative learning analytics, such as periodic private data checkpoint validation, should be further investigated to determine the effectiveness in mitigating overfitting to synthetic data. As research into synthetic educational data is still in the early stages, further investigation into synthetic data generation methods is required, in particular the use of Generative Adversarial Networks that have produced promising results in other fields for large private datasets [35]. However, it is important to take into consideration the nature of the learning systems when designing synthetic generation models and this will require further investigation to identify suitable methods. Finally, methods of synthetic data sharing that have been shown to be effective through empirical evaluation could be integrated into existing learning analytics platforms at various levels to systematically support the sharing of data. This could improve access to critical data for research by third parties for the betterment of education and further development of the educational data mining and learning analytics research communities.

REFERENCES

[1] M. Khalil and M. Ebner, "De-identification in learning analytics," *J. Learn. Anal.*, vol. 3, no. 1, pp. 129–138, Apr. 2016.

[2] A. Pardo and G. Siemens, "Ethical and privacy principles for learning analytics," *Brit. J. Educ. Technol.*, vol. 45, no. 3, pp. 438–450, May 2014.

[3] H. Drachsler and W. Greller, "Privacy and analytics: It's a DELICATE issue a checklist for trusted learning analytics," in *Proc. 6th Int. Conf. Learn. Anal. Knowl.*, 2016, pp. 89–98.

[4] C. M. Steiner, M. D. Kickmeier-Rust, and D. Albert, "LEA in private: A privacy and data protection framework for a learning analytics toolbox," *J. Learn. Anal.*, vol. 3, no. 1, pp. 66–90, Apr. 2016.

[5] M. E. Gursoy, A. Inan, M. E. Nergiz, and Y. Saygin, "Privacy-preserving learning analytics: Challenges and techniques," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 68–81, Jan. 2017.

[6] B. Flanagan and H. Ogata, "Learning analytics platform in higher education in Japan," *Knowl. Manage. E-Learn., Int. J.*, vol. 10, no. 4, pp. 469–484, 2018.

[7] G. Siemens, "Learning analytics: The emergence of a discipline," *Amer. Behav. Scientist*, vol. 57, no. 10, pp. 1380–1400, 2013.

[8] T. Hoel and W. Chen, "Privacy-driven design of learning analytics applications—Exploring the design space of solutions for data sharing and interoperability," *J. Learn. Anal.*, vol. 3, no. 1, pp. 139–158, Apr. 2016.

[9] D. Ifenthaler and C. Schumacher, "Student perceptions of privacy principles for learning analytics," *Educ. Technol. Res. Develop.*, vol. 64, no. 5, pp. 923–938, Oct. 2016.

[10] T. Hoel, D. Griffiths, and W. Chen, "The influence of data protection and privacy frameworks on the design of learning analytics systems," in *Proc. 7th Int. Learn. Anal. Knowl. Conf.*, Mar. 2017, pp. 243–252.

[11] Y. S. Tsai and D. Gasevic, "Learning analytics in higher education—Challenges and policies: A review of eight learning analytics policies," in *Proc. 7th Int. Learn. Anal. Knowl. Conf.*, 2017, pp. 233–242.

[12] C. Fischer, Z. A. Pardos, R. S. Baker, J. J. Williams, P. Smyth, R. Yu, S. Slater, R. Baker, and M. Warschauer, "Mining big data in education: Affordances and challenges," *Rev. Res. Educ.*, vol. 44, no. 1, pp. 130–160, Mar. 2020.

[13] A. Hershkovitz, S. Knight, J. Jovanovic, S. Dawson, and D. Gasevic, "Research with simulated data," *J. Learn. Anal.*, vol. 4, no. 1, pp. 1–2, 2017.

[14] D. Bonnéry, Y. Feng, A. K. Henneberger, T. L. Johnson, M. Lachowicz, B. A. Rose, T. Shaw, L. M. Stapleton, M. E. Woolley, and Y. Zheng, "The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data," *J. Res. Educ. Effectiveness*, vol. 12, no. 4, pp. 616–647, Oct. 2019.

[15] R. Ferguson, T. Hoel, M. Scheffel, and H. Drachsler, "Guest editorial: Ethics and privacy in learning analytics," *J. Learn. Anal.*, vol. 3, no. 1, pp. 5–15, Apr. 2016.

[16] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010.

[17] M. Agaoglu, "Predicting instructor performance using data mining techniques in higher education," *IEEE Access*, vol. 4, pp. 2379–2387, 2016, doi: 10.1109/ACCESS.2016.2568756.

[18] K. T. Chui, R. W. Liu, M. Zhao, and P. O. D. Pablos, "Predicting students' performance with school and family tutoring using generative adversarial network-based deep support vector machine," *IEEE Access*, vol. 8, pp. 86745–86752, 2020, doi: 10.1109/ACCESS.2020.2992869.

[19] S. D. A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita, and N. A. M. Ghani, "Multiclass prediction model for student grade prediction using machine learning," *IEEE Access*, vol. 9, pp. 95608–95621, 2021, doi: 10.1109/ACCESS.2021.3093563.

[20] A. Alshanqiti and A. Namoun, "Predicting student performance and its influential factors using hybrid regression and multi-label classification," *IEEE Access*, vol. 8, pp. 203827–203844, 2020, doi: 10.1109/ACCESS.2020.3036572.

[21] P. M. Moreno-Marcos, T.-C. Pong, P. J. Munoz-Merino, and C. D. Kloos, "Analysis of the factors influencing learners' performance prediction with learning analytics," *IEEE Access*, vol. 8, pp. 5264–5282, 2020, doi: 10.1109/ACCESS.2019.2963503.

[22] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan, "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021, doi: 10.1109/ACCESS.2021.3049446.

[23] E. Popescu and F. Leon, "Predicting academic performance based on learner traces in a social learning environment," *IEEE Access*, vol. 6, pp. 72774–72785, 2018, doi: 10.1109/ACCESS.2018.2882297.

[24] L. Zhao, K. Chen, J. Song, X. Zhu, J. Sun, B. Caulfield, and B. M. Namee, "Academic performance prediction based on multisource, multifeature behavioral data," *IEEE Access*, vol. 9, pp. 5453–5465, 2021, doi: 10.1109/ACCESS.2020.3002791.

[25] R. Alamri and B. Alharbi, "Explainable student performance prediction models: A systematic review," *IEEE Access*, vol. 9, pp. 33132–33143, 2021, doi: 10.1109/ACCESS.2021.3061368.

[26] K. VanLehn, S. Ohlsson, and R. Nason, "Applications of simulated students: An exploration," *J. Artif. Intell. Educ.*, vol. 5, p. 135, Feb. 1994.

[27] G. McCalla and J. Champaign, "Simulated learners," *IEEE Intell. Syst.*, vol. 28, no. 4, pp. 67–71, Jul. 2013.

[28] G. McCalla and J. Champaign, "AIED 2013 simulated learners workshop," in *Proc. Int. Conf. Artif. Intell. Educ.* Berlin, Germany: Springer, 2013, pp. 954–955, doi: 10.1007/978-3-642-39112-5_164.

[29] J. Niznan, J. Papousek, and R. Pelánek, "Exploring the role of small differences in predictive accuracy using simulated data," in *Proc. AIED Workshops*, 2015, pp. 1–10.

[30] Z. A. Pardos and M. V. Yudelson, "Towards moment of learning accuracy," in *Proc. AIED Workshops Volume*, vol. 4, 2013, pp. 65–84.

[31] R. Azoulay, E. David, M. Avigal, and D. Hutzler, "Adaptive task selection in automated educational software: A comparative study," in *Intelligent Systems and Learning Data Analytics in Online Education*, S. Caballé, S. N. Demetriadis, E. Gómez-Sánchez, P. M. Papadopoulos, and A. Weinberger, Eds. New York, NY, USA: Academic Press, 2021, pp. 179–204, doi: 10.1016/B978-0-12-823410-5.00008-5.

[32] D. B. Rubin, "Statistical disclosure limitation," *J. Off. Statist.*, vol. 9, no. 2, pp. 461–468, 1993.

[33] N. Antulov-Fantulin, M. Bošnjak, V. Zlatić, M. Grčar, and T. Šmuc, "Synthetic sequence generator for recommender systems–memory biased random walk on a sequence multilayer network," in *Proc. Int. Conf. Discovery Sci.* Cham, Switzerland: Springer, 2014, pp. 25–36, doi: 10.1007/978-3-319-11812-3_3.

[34] B. Jelic, R. Grbic, M. Vranjes, and D. Mijic, "Can we replace real-world with synthetic data in deep learning-based ADAS algorithm development?" *IEEE Consum. Electron. Mag.*, early access, May 26, 2021, doi: 10.1109/MCE.2021.3083206.

[35] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, "Synthetic data in machine learning for medicine and healthcare," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 493–497, Jun. 2021.

[36] K. E. Emam, "Seven ways to evaluate the utility of synthetic data," *IEEE Secur. Privacy*, vol. 18, no. 4, pp. 56–59, Jul. 2020.

[37] H. Ping, J. Stoyanovich, and B. Howe, "DataSynthesizer: Privacy-preserving synthetic datasets," in *Proc. 29th Int. Conf. Sci. Stat. Database Manage.*, Jun. 2017, pp. 1–5.

[38] J. Dahmen and D. Cook, "SynSys: A synthetic data generation system for healthcare applications," *Sensors*, vol. 19, no. 5, p. 1181, Mar. 2019.

[39] A. M. Berg, S. T. Mol, G. Kismihók, and N. Sclater, "The role of a reference synthetic data generator within the field of learning analytics," *J. Learn. Anal.*, vol. 3, no. 1, pp. 107–128, Apr. 2016.

[40] M. Dorodchi, E. Al-Hossami, A. Benedict, and E. Demeter, "Using synthetic data generators to promote open science in higher education learning analytics," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 4672–4675.

[41] A. Peña-Ayala, "Learning analytics: A glance of evolution, status, and trends according to a proposed taxonomy," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 3, p. e1243, May 2018.

[42] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 3, p. e1355, May 2020.

[43] S. B. Shum, M. Hawksey, R. S. J. D. Baker, N. Jeffery, J. T. Behrens, and R. Pea, "Educational data scientists: A scarce breed," in *Proc. 3rd Int. Conf. Learn. Anal. Knowl.*, 2013, pp. 278–281.

[44] J. Gardner, C. Brooks, J. M. Andres, and R. S. Baker, "MORF: A framework for predictive modeling and replication at scale with privacy-restricted MOOC data," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 3235–3244.

[45] H. Ogata, M. Oi, K. Mohri, F. Okubo, A. Shimada, M. Yamada, J. Wang, and S. Hirokawa, "Learning analytics for e-book-based educational big data in higher education," in *Smart Sensors at the IoT Frontier*. Cham, Switzerland: Springer, 2017, pp. 327–350, doi: 10.1007/978-3-319-55345-0_13.

[46] M. Oi, F. Okubo, A. Shimada, C. Yin, and H. Ogata, "Analysis of preview and review patterns in undergraduates'e-book logs," in *Proc. 23rd Int. Conf. Comput. Educ.*, 2015, pp. 166–171.

[47] R. Majumdar, A. Akçapınar, G. Akçapınar, H. Ogata, and B. Flanagan, "LAView: Learning analytics dashboard towards evidence-based education," in *Proc. Companion 9th Int. Conf. Learn. Anal. Knowl., Soc. Learn. Anal. Res. (SoLAR)*, 2019, pp. 1–7.

[48] B. O. Ngoko, H. Sugihara, and T. Funaki, "Synthetic generation of high temporal resolution solar radiation data using Markov models," *Sol. Energy*, vol. 103, pp. 160–170, May 2014.

[49] J. Schreiber, "Pomegranate: Fast and flexible probabilistic modeling in Python," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 5992–5997, 2017.

[50] G. Akçapınar, M. N. Hasnine, R. Majumdar, B. Flanagan, and H. Ogata, "Developing an early-warning system for spotting at-risk students by using eBook interaction logs," *Smart Learn. Environments*, vol. 6, no. 1, p. 4, Dec. 2019.

[51] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[52] F. Okubo, T. Yamashita, A. Shimada, and H. Ogata, "A neural network approach for students' performance prediction," in *Proc. 7th Int. Learn. Anal. Knowl. Conf.*, Mar. 2017, pp. 598–599.

[53] D. Gašević, S. Dawson, T. Rogers, and D. Gasevic, "Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success," *Internet Higher Educ.*, vol. 28, pp. 68–84, Jan. 2016.

[54] R. Pelánek, J. Rihák, and J. Papoušek, "Impact of data collection on interpretation and evaluation of student models," in *Proc. 6th Int. Conf. Learn. Anal. Knowl.*, 2016, pp. 40–47.

[55] H. Li, R. Majumdar, M.-R.-A. Chen, and H. Ogata, "Goal-oriented active learning (GOAL) system to promote reading engagement, self-directed learning behavior, and motivation in extensive reading," *Comput. Educ.*, vol. 171, Oct. 2021, Art. no. 104239.

BRENDAN FLANAGAN (Member, IEEE) received the bachelor's degree from RMIT University and the master's and Ph.D. degrees from the Graduate School of Information Science and Electrical Engineering, Kyushu University. He is currently a Senior Lecturer with the Academic Center for Computing and Media Studies, Kyoto University. His research interests include learning analytics, text mining, machine learning, and language learning.

RWITAJIT MAJUMDAR (Member, IEEE) is currently a Senior Lecturer with the Academic Center for Computing and Media Studies, Kyoto University, associated with the Learning and Educational Technologies Research Unit. He did his doctoral Research with the Indian Institute of Technology Bombay in the Interdisciplinary Program of Educational Technology. His research interests include learning analytics, visual analytics of educational data, and human–computer interactions.

HIROAKI OGATA (Senior Member, IEEE) is currently a Professor with the Academic Center for Computing and Media Studies, the Learning and Educational Technologies Research Unit, and the Graduate School of Informatics, Kyoto University, Japan. He is also leading a research project on development of infrastructure for learning analytics and educational data science. His research interests include learning analytics, evidence-based education, educational data mining, educational data science, computer supported ubiquitous and mobile learning, computer supported collaborative learning (CSCL), computer supported collaborative writing (CSCW), computer assisted language learning (CALL), computer supported social networking (CSSN), knowledge awareness, personalized, and adaptive and smart learning environments.