Evaluating the Impact of Data Augmentation on Predictive Model Performance

Valdemar Švábenský
Kyushu University
Fukuoka, Japan
valdemar.research@gmail.com

Conrad Borchers
Carnegie Mellon University
Pittsburgh, PA, USA
cborcher@cs.cmu.edu

Elizabeth B. Cloude
Michigan State University
Lansing, MI, USA
ecloude94@gmail.com

Atsushi Shimada
Kyushu University
Fukuoka, Japan
atsushi@limu.ait.kyushu-u.ac.jp

Abstract

In supervised machine learning (SML) research, large training datasets are essential for valid results. However, obtaining primary data in learning analytics (LA) is challenging. Data augmentation can address this by expanding and diversifying data, though its use in LA remains underexplored. This paper systematically compares data augmentation techniques and their impact on prediction performance in a typical LA task: prediction of academic outcomes. Augmentation is demonstrated on four SML models, which we successfully replicated from a previous LAK study based on AUC values. Among 21 augmentation techniques, SMOTE-ENN sampling performed the best, improving the average AUC by 0.01 and approximately halving the training time compared to the baseline models. In addition, we compared 99 combinations of chaining 21 techniques, and found minor, although statistically significant, improvements across models when adding noise to SMOTE-ENN (+0.014). Notably, some augmentation techniques significantly lowered predictive performance or increased performance fluctuation related to random chance. This paper's contribution is twofold. Primarily, our empirical findings show that sampling techniques provide the most statistically reliable performance improvements for LA applications of SML, and are computationally more efficient than deep generation methods with complex hyperparameter settings. Second, the LA community may benefit from validating a recent study through independent replication.

CCS Concepts

• **Computing methodologies** → **Machine learning**; *Modeling and simulation*; • **Applied computing** → **Education**.

Keywords

learning analytics, prediction, supervised learning, data generation, synthetic data, replication

LAK 2025, Dublin, Ireland

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706485

ACM Reference Format:
Valdemar Švábenský, Conrad Borchers, Elizabeth B. Cloude, and Atsushi Shimada. 2025. Evaluating the Impact of Data Augmentation on Predictive Model Performance. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706485

1 Introduction

To improve teaching and learning, primary student data (such as data on students' learning progression, behavior, and outcomes) are essential for studying scientific questions in learning analytics (LA) research. High quality and adequate volume of research data are especially important when the research methods involve machine learning (ML), such as predictive modeling, which is extensively used in LA [21, 57]. However, collecting such data is challenging due to several reasons. First, it is time-intensive and costly to capture data from a sufficiently large sample of participants. In addition, the data samples usually represent homogeneous learner populations, meaning that they are bound to a specific context, time, and learning system [5]. As a result, datasets can often be small and have low diversity, limiting predictive model performance and generalizability. Second, ensuring data privacy requires thorough anonymization, and while tools are being developed to help automate this task [61], it is still time-consuming and error-prone. Third, additional ethical and legal issues are associated with collecting, storing, and processing student data [30, 37]. This is especially true for sensitive data collected in schools or other research settings that can personally identify a student. Prominent examples of such data include students' writing style or speech, for example, in discussion forum posts [27], student discourse [15], or written peer feedback on problem-solving [29].

1.1 Data Augmentation

There is opportunity to address the aforementioned challenges of many datasets' limitations in terms of size and diversity by utilizing *data augmentation*: a set of techniques [42] to "increase the volume, quality and diversity of training data" [46], while simultaneously maintaining students' privacy and adhering to ethical standards of research. Data augmentation involves computational methods for sampling new data, transforming existing data, and generating synthetic data [66] (see details in Section 2.2). At their core, these methods enhance the predictive performance of ML models by

leveraging underutilized signals in imbalanced data or rare data patterns. At the same time, they regularize the model through noise addition, akin to other regularization techniques like dropout or weight penalties.

These techniques have proven effective for improving ML model performance and generalizability in data science applications across various domains [12], such as image processing [60], computer vision [31], and healthcare [53]. However, the utility of data augmentation depends on dataset characteristics (e.g., size, structure, sparsity) [7, 60]. In non-LA domains [7, 31, 53, 60], augmented datasets often have properties different from typical LA datasets. Since augmentation is underexplored within LA-relevant datasets, much is still unknown about its effect and influence on ML model performance in LA research. Therefore, its viability must be rigorously evaluated to improve critical LA tasks.

Data augmentation, especially synthetic data generation, is closely related to simulated data. Recently, Käser and Alexandron [35] published "the first systematic literature review on simulated learners", which uncovered challenges and limitations for current simulated LA datasets. First, they argued that "simulated learner models tend to represent only narrow aspects of student learning", as most prior work focused on simulating a small set of cognitive skills [35]. Yet, non-cognitive factors are also essential to learning and influence student outcomes [2]. Thus, learner models should represent multiple facets of learning guided by theoretical perspectives on human knowledge and skill acquisition. Second, several studies do not provide evidence on the validity of simulated learner models and downstream findings generated from them. More research is needed to address these limitations. This is one of the aims of the current study, which uses data augmentation to improve model predictive validity in a common LA classification task.

1.2 Research Reproducibility and Replicability

An additional aspect that improves the validity of learner models and the ability to draw generalizable conclusions is the successful reproduction and replication of research findings. According to the Association for Computing Machinery (ACM), *reproducibility* of research means that a result from a computational experiment can be independently obtained by a different team (other than the original authors) using data and artifacts from the original author team [1]. This is a key property of good science, as it ensures credible results and enables their broader applicability. A term related to reproducibility is *replicability*, which means that a result can be obtained by a different team using different artifacts [1]. Currently, very few published LA studies have been replicated, which we describe more thoroughly in Section 2.1. Moreover, AI research as a whole (which includes machine learning in LA), also faces a reproducibility crisis [28].

1.3 Goals and Scope of This Paper

Our paper uniquely addresses both replication and data augmentation in the LA context. Specifically, we augment existing student data selected from studies recently published at the LAK conference. Since these studies have already passed the peer review process, the existing models should have higher validity than if we developed new models from scratch. To narrow our focus, we choose the context of predictive ML modeling, which is central to LA research [21, 57]. ML is a leading topic especially at conferences like LAK due to its implications for building personalized and adaptive systems that guide just-in-time interventions, such as dropout prediction [57], knowledge tracing [3], affect detection [33], and the evaluation of predictive analytics interventions in the classroom [38]. However, valid ML research requires robust models and generalizable findings [5], which can be supported by replication and data augmentation.

Focusing on research that reported predictive ML models, we first attempt to *replicate* these models and the corresponding findings. Second, to evaluate the utility of data augmentation, we *re-train* these models with a mixed dataset consisting of original and augmented data. We hypothesize that the re-trained ML models will demonstrate better performance than the original ML models due to better quality and more diverse data for training.

Within this scope, our paper investigates the following research questions (RQs):

RQ1 *To what extent can we replicate the analyses and results from a selected previous learning analytics study?*

RQ2 *When data augmentation techniques are applied to the original training data, which of these techniques, and to what extent, improve model performance?*

1.4 Contributions

Our paper is original in tackling both the issues of replication and data augmentation in parallel. Combining replication and augmentation is valuable for two reasons. First, modeling methods were validated through prior peer-review, ensuring they meet LAK's methodological community standards (e.g., cross-validation). Second, replicating published models establishes a competitive baseline for demonstrating reliable performance improvements through augmentation.

Our paper's first contribution is replicating an LA study that predicts long-term academic outcomes using ML models with features grounded in a learning theory. Second, we use the replicated models as a case study to demonstrate and evaluate whether augmentation techniques improve prediction, and in which cases. We measure changes in model performance to systematically compare whether different augmentation techniques yield statistically significant improvements. Third, we offer a methodological contribution in showing how to evaluate augmentation techniques and offer practical recommendations and lessons learned. All code is available for other researchers to adopt or adapt (see Section 5). This enables researchers to augment their own datasets and study the significance of prediction performance improvements in relationship to a suite of augmentation techniques for their own educational models.

From a high-level perspective of the four phases of the LA cycle [14], our research focuses on improving *data* to build more robust, valid, and generalizable models during the *metrics* phase. As a result, improving ML models holds implications for yielding more accurate educational *interventions* to better support *learners*. Lastly, it contributes to creating more effective personalized systems that can better support teaching and learning practices.

2 Related Work and Novelty of This Paper

We review prior work for each RQ: Section 2.1 focuses on replication, and Section 2.2 on data augmentation. In both sections, the last paragraph highlights the novelty of our paper. In addition, Section 2.3 introduces theory that grounded our approach – another aspect that has been insufficiently covered in literature [35] and is addressed by our paper.

2.1 RQ1: Research Reproducibility and Replicability in LA

In the ACM Digital Library, which indexes all past 14 years of LAK conference publications, we searched for terms related to reproducibility and replicability (search query: `reproduc* OR replica*`) in the titles, abstracts, or keywords of the published articles. Constraining the search to the most recent five years of LAK (2020–2024) yielded nine papers, out of which seven performed reproduction or replication as defined in Section 1.2. Within this five-year search period, a total of 379 full and short papers were published at LAK, which means that only 1.8% included an element of reproducing/replicating previous research. Moreover, out of these 1.8% that replicated past research results, 0% performed data augmentation with the aim of improving predictive performance, as explored in our present study.

The finding that only seven of the 379 (1.8%) LAK papers in the past five years reproduced/replicated a past study can partially be attributed to the lack of reproducibility described by Haim et al. [23]. They estimated the reproducibility of all LAK 2021 and 2022 papers, investigating whether the papers' data and supplementary materials (e.g., code) were documented well, such that another researcher could (with reasonable effort) reproduce the findings. Only 5% of papers made their raw dataset available, and in another 2%, data could be requested from the authors, effectively making at least 93% of LAK papers *not reproducible*. In the related field of educational data mining (EDM), the ratios of data sharing practices were slightly higher, but still low: 15% available and 5% on request [22], making 80% of papers unreproducible.

Ultimately, Haim et al. [23] marked all LAK 2021–2022 papers as *not reproducible* within the time slot they allocated for each paper, but "estimated that the 2% of papers that contain both the raw dataset and source [code] were likely to be reproducible". This is a surprisingly low number, given that datasets and computational methods are central to the progress of LA research. Despite efforts that strongly recommend and promote the sharing of data and code among researchers [36], it is still not a prevalent practice. Since reproducing research is so rare, this creates a gap in the LA literature, which may lead to questioning the validity and generalizability of some LA results. Our paper contributes towards addressing this gap by validating a recently published study through independent replication.

2.2 RQ2: Data Augmentation and Its Use in LA

In line with our goals (see Section 1.3), we surveyed data augmentation approaches used for improving predictive performance of ML models. We distinguish between *sampling*, *perturbation*, and *generation*. Within these categories, the individual techniques are often *symbolic* (using statistical models) or *rule-based* (using programs or templates) [42].

Sampling methods create additional data points by resampling from an existing dataset. Common approaches in classification contexts include oversampling, where minority class instances are duplicated, and undersampling, where majority class instances are reduced. A prominent example is the Synthetic Minority Oversampling Technique (SMOTE), which creates synthetic samples for the minority class by interpolating between existing observations [11]. Adaptive Synthetic Sampling (ADASYN) improves oversampling by creating more samples in regions with sparser data distribution, thereby adapting to the difficulty of learning in different regions of the feature space [25]. Both SMOTE and ADASYN slightly improved the accuracy of student performance prediction in a prior LA study [4].

Perturbation methods (also called *data transformation* [46]) involve making small, controlled changes to the original data. They modify the existing training samples or the derived features to increase variance in the feature space during model training. Representative techniques such as noise injection add random values to existing samples, enhancing the models' generalization by preventing overfitting to the original data [59]. Wei and Zou [65] proposed four augmentation techniques using random insertion, swapping, and deletion of words in text classification, boosting test set performance on small datasets. To the best of our knowledge, no past work has systematically investigated perturbation in LA.

Generation methods (also called *data synthesis* [46]) for creating new data have diverse LA applications: from data anonymization, through benchmarking [9], to improving ML models [69]. Various learning processes and outcomes have been studied using synthetic data [9, 68] such as tabular data [39], interaction data [20], simulated learner models [35], or biased data [32]. GPT-4 produced a synthetic dataset of a student-tutor conversation [63]. Next, generative adversarial networks (GANs) are especially useful for imbalanced data as they can oversample the minority class [10]. Zhang et al. [69] used GANs to augment log data from tutoring systems, improving the reliability of assessing student knowledge estimates using the augmented samples. This reliability was not improved when GPT-4 was used.

In past research, the validity of synthetic data has often not been systematically studied [35], raising the question of whether synthetic data accurately represent real learner data. This is crucial because valid measures for assessment and outcome prediction are central LA goals [17]. A common approach for determining validity is *predictive validity*: a measure is valid if it accurately predicts outcomes of interest [38, Chapter 3]. For example, standardized test scores are a valid representation of academic aptitude if they can predict students' future academic success. In line with this definition, our study addresses the lack of prior validation of synthetic data and augmentation methods. We compare the performance of models predicting long-term academic outcomes, with and without augmentation methods applied.

2.3 Learning Theory and Theoretical Grounding for RQ1 and RQ2

Only 3% of simulated learner models in LA and related fields were anchored in theories of human learning [35]. Moreover, simulated learning often overemphasizes cognitive aspects [41], omitting non-cognitive facets of learning such as affective states (e.g., confusion, boredom, engaged concentration). However, to simulate realistic and holistic learner models, it is essential to consider these non-cognitive facets as well, since they influence outcomes and achievement [18].

We ground our RQs in the *Model of Affective Dynamics* [18] (MAD), a prominent contemporary theory that describes human learning. It explains how cognition, knowledge, and affect interact as learners engage with digital learning systems. As learners assimilate new information into their existing knowledge, they often encounter impasses – discrepancies between new information and prior understanding – leading to confusion. Whether this confusion is resolved determines the transition into different affective states, some of which hinder learning. If learners cannot overcome the impasse, confusion may escalate into frustration, and if it persists, further into boredom and disengagement. Conversely, if the impasse is resolved, learners return to a state of engaged concentration, which is conducive to learning.

Empirical evidence shows that learners' affective states significantly influence their long-term academic outcomes, such as college enrollment [51]. Outcomes are positively correlated with engaged concentration [2] and negatively with boredom and disengagement (e.g., gaming the system) [45]. However, Karumbaiah et al. [34]'s findings contradict MAD, implying that individual factors (such as the learner's knowledge) influence how affective states interact with cognition [13], impacting outcomes differently. Specifically, the impact of boredom, confusion, and frustration remains inconclusive: some studies suggest these states are beneficial, while others find them harmful [6, 19, 48, 54]. These inconsistencies could be explained by insufficient or low-quality data, which our study attempts to address.

3 Research Methods

The goal of the present research is to replicate an LA study that involves predicting an academic outcome, and subsequently investigate whether data augmentation boosts model performance on the same task. Initially, we defined criteria for selecting a study for replication (see Section 3.1). Next, Section 3.2 describes the criteria application and relevant aspects of the selected LA study. Section 3.3 details our methods for replicating the study (RQ1). Finally, Section 3.4 proposes and justifies the methods for data augmentation (RQ2). Figure 1 illustrates our overall approach.

3.1 Defining and Justifying the Criteria for Selecting a Paper to Replicate

To choose a study for replication (RQ1), the research team discussed criteria that a paper should satisfy to attempt replication. When it comes to research data and methods documentation, we drew inspiration from Haim et al. [23], who systematically studied replication in LA. After several iterations and refinement, we defined that the study must:

(1) Be a full or short paper published at the *LAK conference within the past five years* (i.e., from 2020 to 2024).
(2) Have *no overlap of co-authors* with the team of our paper, to perform true replication by a different team.
(3) Have research data that are:
   (a) *Open*, which we define – in accordance with Haim et al. [23] – as available either through a free online download, or on request with a latency of less than one month for anyone in the broader research community.
   (b) *In English language* for all of the associated results, as there are some LAK 2020–2024 papers with data in Portuguese or Chinese, enabling the data reuse and interpretation only to researchers fluent in these languages.
   (c) *Not simulated/synthetic*, since one of the goals of our paper is to determine improvements of augmenting real-world learner data with synthetic or otherwise modified data (RQ2).
(4) Have research methods that are *documented*, which we define – in line with Haim et al. [23] – as allowing us to reproduce or replicate the original study's results (e.g., key outcome measures) within reasonable time.
(5) Have an analytical research objective that involves *prediction by using supervised machine learning* (SML), a central application area of LA [21]. This is because our RQ2 aims to evaluate augmentation techniques, which are typically used for improving the prediction of outcomes of interest through augmented predictors [4, 10, 59, 65].
   (a) The relationship between the outcome variable and the predictors must be grounded in a learning theory, involving learning process features and academic outcomes derived from educational data.
   (b) The SML models must use a test set for evaluation. This allows for reliably estimating predictive performance improvements achieved by data augmentation applied to the training data and tested on the separate test set.
   (c) The modeling methods must follow educational data science research standards [38, 55], including cross-validation and a comparison of several prediction schemes with different architectural complexity. This ensures any improvements to model performance can be attributed to augmentation, not lack of methodological rigor.

3.2 Selected Paper for Replication

At the beginning, 379 candidate papers satisfied criterion (1), as also shown in Section 2.1. Applying criterion (2) left us with 371 papers. Out of these, 36 had open research data based on criterion (3a), which is 9.7%. This ratio of data sharing is similar to, though slightly higher than, the 7% reported by Haim et al. [23]. Applying the rest of criterion (3), we were left with 31 papers. Out of these, 18 papers partially satisfied (4), although some rather minimally, and two papers satisfied (5). For one of the two papers, upon more thorough inspection we realized that we needed additional data from the authors to replicate the findings, but they did not reply to our email and thus the paper was discarded.

We ultimately selected the study by Zambrano and Baker [67], which investigated whether students' topic-level knowledge in

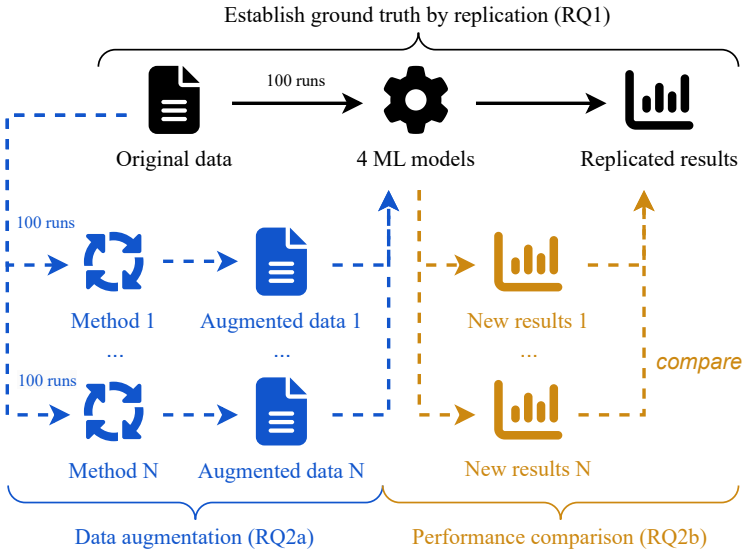

Figure 1: Conceptual illustration of the research methods and steps, along with the relationship between the two RQs.

mathematics along with learning behaviors could predict two outcomes: *college enrollment* and *STEM career choice*. This study was grounded in a theoretical framework of human learning [18] and discussed the empirical evidence that justified the selection of features used in the ML models. The study topic is directly related to the key goals of LA, which center around generating insights to support teaching, learning, and educational management [38].

3.2.1 Dataset. The study used a public dataset [49] from ASSISTments: a digital learning platform designed to improve mathematical skills [26]. There were 1,709 students from four middle schools in the USA, who practiced tasks grouped into 12 topic areas [67, Table 1], such as algebra, geometry, and functions. This dataset tracked students longitudinally: starting from their interactions with ASSISTments and behaviors during learning activities, to their college enrollment and their first job (STEM vs. non-STEM). We obtained free access to the dataset after filling in the request form [49].

3.2.2 Target Variable. We decided to focus only on the *enrollment* outcome and not *STEM career*. This is because for predicting *enrollment*, the paper used the full dataset of 1,709 students, but for *STEM career*, only about a third of the dataset (591 students) included a coded outcome. This small sample size would make it difficult to establish significantly different ML model performance results between baseline and augmented models in our RQ2. Specifically, based on an ad-hoc power analysis, we determined that an AUC difference of 0.05 can be reliably detected using DeLong's test [16, 64] by a sample of 1,709 students (> 90%) but not 591 students (~40%). Even though some data augmentation techniques help deal with small sample sizes, the small sample size could bias the initial replicated model for which we need to compare the augmented dataset against. Lastly, the class distribution of the selected target variable is slightly imbalanced: 1,097 students (64%) enrolled in college (class 1), while the remaining 612 (36%) did not (class 0).

3.2.3 Predictor Variables. To predict enrollment, two types of features were used: *cognitive* and *affective*, which are both theoretically and empirically supported as influencing student outcomes [2, 18]. The *cognitive* features were students' skill estimates within the 12 topics in mathematics [67, Table 2]. To estimate learners' mastery for each topic, Bayesian Knowledge Tracing (BKT) was used [52]. BKT is a widely recognized technique for tracing learners' knowledge and predicting their knowledge states using data from multiple learning tasks or assessments [3]. Employing a commonly-used algorithm that demonstrated satisfactory performance across a wide range of dataset sizes and learning contexts [62], BKT estimates a learner's current state of knowledge for a specific topic over time, using performance data generated during the interaction with a digital learning platform. Specifically, BKT estimates the probability (a real number between 0 and 1) of a learner correctly applying an underlying skill (e.g., division) to a problem without any instructional support. The *affective* features were six measures of students' emotions or behavior: boredom, concentration, confusion, frustration, off-task behavior, and gaming the system. These features align with the model of affective dynamics [18] and have been commonly used to understand and predict student outcomes [33]. Due to the space limitations of this paper, further details are available in our code documentation (see link in Section 5).

3.2.4 Prediction Models. The original study [67] employed SML using four different binary classification algorithms: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Multi-Layer Perceptron (MLP). Their implementations come from the open-source Python library `scikit-learn` [50], and the original study [67] used them with default hyperparameters. The metric used to evaluate the models' classification performance is the Area Under the Receiver Operating Curve (AUC) [24], which has been commonly used at LAK [32, 39]. It describes the probability of a classifier correctly distinguishing a positive and negative outcome observation. For each of the four models, the original

reported AUC was approximately 0.69 [67, Table 6], which is a somewhat satisfactory performance.

3.3 RQ1: Paper Replication Process

Two co-authors of this paper used the ASSISTments dataset [49] and the descriptions in the original study [67] to implement Python code for replicating the predictive models. A third co-author performed independent testing and code review. Since the original paper does not have the source code available, we implemented ours from scratch. Nevertheless, we later received a private copy of the original code after contacting the authors, so we used this original code to validate our approach. Our version of the implementation is publicly available (see Section 5).

The original study [67, Section 3] described most modeling steps, eliminating the need to make arbitrary decisions on our side. After data preprocessing, we verified that the cognitive features matched the original paper within a small error margin. The overall mean absolute difference between the feature values in the paper [67, Table 2] and ours was only 0.013. Since the standard deviation of the expected parameter distribution was 0.135, and 0.134 for our computed values, the difference of 0.013 is negligible. The affective features were fully equivalent to the original paper.

The original study used only two classification models (LR and SVM) that are deterministic, meaning that results and model fits are equivalent for each training run of the model. We empirically confirmed that both indeed behave so. In contrast, the other two models (RF and MLP) incorporate randomness, and the original study did not use a fixed seed. This would make it difficult to estimate whether our model performances are better or worse by chance, and to what extent they fluctuate by chance. Therefore, to account for the performance variations, we trained each non-deterministic model 100 times. This iteration count was chosen given the computation time and resource constraints. Subsequently, to ensure a fair performance comparison, we report the average AUCs across the 100 runs (which will also be our baseline for comparing augmentation results). The relevant metric that is averaged across runs is the average AUC across folds, following Zambrano and Baker [67]. Moreover, to ensure the reproducibility of our results, we set fixed, arbitrary random seeds, which differ for each run. Due to using various seeds and iterating 100 times, this setting ensures that the results of the non-deterministic models are more stable from a random sampling perspective.

3.4 RQ2: Data Augmentation Process

After the replication, we assessed numerous classical and contemporary augmentation techniques representing different architectural complexity, ensuring reproducibility by using off-the-shelf libraries. We compared each technique's results to a baseline without augmentation to determine which is the most effective for the given prediction task. Based on related work reviewed in Section 2.2, we implemented 21 techniques grouped into 3 categories described in Table 1.

We tested all 21 techniques for each of the four models. For each technique, we augmented the input dataset and executed the same modeling code as in RQ1, with a single modification: removing the forward feature selection (FFS) present in the original study and during replication. The reason was that during test runs, it more than quadrupled the execution time, making the convergence of generation methods almost infeasible. Other than that, using the same approach enabled us to assess if augmentation is effective for some model architectures but not others.

Due to numerous execution combinations (see below), we employed three hardware configurations in parallel:

- Laptop: 13th Gen Intel(R) Core(TM) i7-1360P, 2.20 GHz. 16 GB RAM.
- Workstation: 13th Gen Intel(R) Core(TM) i9-13900K, 3.00 GHz. 128 GB RAM.
- Cluster: Per each task, 1 computing server node out of 150 was used, each with 4 CPU cores and 128 GB RAM.

First, we executed each of the 21 augmentation techniques separately. Most of their hyperparameters were set to default (see the code for details). For all three deep generation techniques, we set the hyperparameters to equal values to achieve fair comparison: latent dimension = number of epochs = 100, number of batches = 64. Ideally, we would have used 1000 epochs, but even with the 100 on the stronger hardware, the runtime still takes several hours (see Table 2).

Second, we combined pairs of augmentation techniques by sequentially chaining their execution, which has improved prediction results in recent SML research [7, 58]. For the 21 techniques, assuming it does not make sense to apply the same technique twice, there are $21 \times 20 = 420$ possibilities to vary their ordering. However, evaluating 420 chained techniques for each model would not be feasible given long runtimes. Fortunately, the literature established precedents for preferring certain combinations: chaining sampling and then generation [58], or chaining perturbation and then sampling [43]. We also chained perturbation and then generation. For the chaining involving generation, we only used the GAN technique, as it is the fastest. Based on Table 1, this gives us $9 \times 1 + 9 \times 9 + 9 \times 1 = 99$ combinations to evaluate, each executed again in 100 iterations for each model. To the best of our knowledge, no LA paper has attempted this kind of composite and systematic evaluation of augmentation techniques.

The number of data points affected by augmentation differs for each category of techniques. For sampling, the minority class data points were boosted to achieve a number of observations equal to the majority class. For perturbation, no new observations were generated and all existing data points were affected by the transformations. For deep generation methods, the size of the dataset was doubled.

Finally, after obtaining the results of all augmented models, we assessed which augmentation techniques significantly improved performance. DeLong's test [16] is the most common method for computing significance between two AUC ROC curves, but it requires predicted class probabilities of a specific model run. Here, we aim to adjust for performance differences that are by chance across several runs, either through specific data points in small samples or random initialization variation in model weights (i.e., seed variation). Therefore, we bootstrapped AUC values within augmentation and baseline methods, meaning we sampled data with replacement and computed the AUC of each sample to induce a distribution of AUC values. These distributions are used to compute empirical

Table 1: Overview of 3 categories of data augmentation methods (we used 21 techniques total) and their LA applications.

Aug. method	**Description and examples** (see Section 2.2 for details)	**References**
Sampling (9 techniques)	Oversampling duplicates minority class instances; undersampling reduces majority class instances. SMOTE generates synthetic minority samples by interpolation, while ADASYN focuses on sparse regions for more targeted oversampling.	[4, 11, 25]
Perturbation (9 techniques)	Small controlled changes to data, like noise injection, to increase variance and prevent overfitting. Wei and Zou [65] proposed random insertion, swapping, and deletion for text data, which improved performance on small datasets.	[59, 65]
Generation (3 techniques)	Generative Adversarial Networks (GANs) create synthetic data by learning the sample distribution, which is particularly useful for imbalanced datasets. GANs have been used in LA to improve knowledge assessment.	[9, 10, 69]

distributions of AUC values per model architecture and augmentation method. The empirical distributions are then compared using a z-transformation to compute critical quantiles and corresponding p-values representing whether predictive methods produce AUC distributions that are significantly different from one another. Given the large number of models compared using p-values, we adjusted α for multiple comparisons using the Benjamini-Hochberg correction in R [44], setting the false discovery rate to the conventional 0.05 threshold [8].

4 Results and Their Discussion

Table 2 provides all results for both RQs. We discuss them separately below. As a reminder, for each row in Table 2, we ran non-deterministic models 100 times to mitigate the effects of randomness. We report the mean and standard deviation of the performance measures across the 100 runs, since it is a common practice in LA research [67, 68].

4.1 RQ1: Paper Replication

Based on the first two rows of Table 2, we replicated similar model performances as in the original paper [67]. The total mean AUC difference between the original and our models is 0.036, which is minor. Moreover, the difference between the original best model and our best model (both LR) is just 0.007. Zambrano and Baker's [67] results are within the 95% confidence interval of our results, which can be considered a successful replication within the uncertainty of seed variation not accounted for in the original paper. Another evidence indicating close replication is that the relative order of the models' performance is preserved with respect to the original study (LR > RF > SVM > MLP). It is interesting that in both original and our study, the simplest model architecture performed the best, while the most complex model performed the worst. The observed performance differences can be attributed to the random variation caused by the unknown seed in the original paper. Seed variation is a common factor preventing exact reproducibility at LAK [23].

Additionally, our results are consistent with previous research, which has demonstrated that both cognitive and affective factors are predictive of college enrollment [51, 56]. These findings also reinforce the central hypothesis of the model of affective dynamics [18], which posits that emotional states and knowledge-related factors (i.e., mathematical skill estimates in this case) over time have a significant impact on long-term academic outcomes.

4.2 RQ2: Data Augmentation

Table 2 also compares our baseline to our augmentation results. Since the replication results are almost exactly the same regardless of whether FFS is used or not, we use the version without FFS as the baseline, because removing FFS substantially reduced the runtime (to about 22% of the original).

4.2.1 Sampling. Overall, *SMOTE-ENN* is the best performing sampling technique, improving the overall mean AUC by 0.01 across model architectures and even reducing the training time to just 55% of the baseline. The improvement is higher for the non-deterministic models, especially MLP (+0.035), where it is also statistically significant. On the contrary, *NearMiss* is the worst technique, reducing all four model performances (significantly in three models) while also prolonging the training by 20%. Among the individual models, LR is barely influenced by any technique (except for the decrease with *NearMiss*). For SVM, *Random Undersampler* (+0.025) and *Random Oversampler* (+0.014) are beneficial.

4.2.2 Perturbation. When applied in isolation, perturbation did not boost model performance. For LR and RF, it had almost no effect. For MLP, the result was almost always worse, often significantly. For SVM, the improvements were smaller than 0.01. Similarly negligible improvement was observed for RF and MLP with *Noise Addition*. However, in all cases, the training time also increased by at least 27%, so perturbation alone was not useful in this prediction task.

4.2.3 Generation. Regarding the deep generation techniques, which are the most complex category, even the best result for each column was always outperformed by at least one simpler sampling technique. In other words, some sampling technique was better than the best generation technique for all four ML model architectures. This is an important finding because the sampling techniques require a fraction of the training time. When compared to perturbation, GAN and CGAN improved the MLP model (+0.014 and +0.022, respectively, compared to the baseline). However, CGAN also notably worsened the LR model, showing that not all techniques are suitable for all model architectures.

4.2.4 Technique Combinations. For the combined augmentation techniques, since the number of combinations is so large, we only report the best combinations per category. The precedents established in the literature (sampling and then generation [58], or perturbation and then sampling [43]) yielded only a microscopic improvement over a single technique (AUC increase of less than 0.01). Specifically,

Table 2: AUC results: mean (and SD in parentheses) across the 100 runs. For each augmentation category, the best result in each column is highlighted in green (in case of tied mean AUCs, the lower SD wins, the other tied results are light green). The worst result in each column is red (in case of tied mean AUCs, the higher SD wins, the other tied results are light red). Next, the bold entries are better than the baseline by at least 0.01. Entries prefixed with an asterisk (*) denote that the AUC is statistically significantly different from the baseline after adjusting for multiple testing. The symbols in the Runtime column refer to the hardware used (see Section 3.4).

Augment. category	Method (FFS = forward feature selection)	Deterministic models LR	Deterministic models SVM	Non-deterministic models RF	Non-deterministic models MLP	Overall mean	Runtime (hh:mm:ss)
None	Original (with FFS) [67]	0.693 (0.050)	0.691 (0.046)	0.692 (0.037)	0.686 (0.041)	0.691 (0.044)	N/A
None	Replication (with FFS)	0.686 (0.047)	0.652 (0.046)	0.654 (0.041)	0.628 (0.032)	0.655 (0.041)	03:45:16
None	Comparison baseline	0.686 (0.047)	0.652 (0.046)	0.654 (0.042)	0.627 (0.032)	0.655 (0.042)	00:48:54
Sampling	SMOTE Standard	0.684 (0.050)	0.659 (0.049)	0.646 (0.046)	0.611 (0.038)	0.650 (0.046)	00:51:43
Sampling	ADASYN	0.688 (0.048)	**0.663** (0.040)	0.638 (0.042)	0.608 (0.033)	0.649 (0.041)	00:43:23
Sampling	BorderlineSMOTE	0.687 (0.048)	**0.663** (0.041)	0.640 (0.040)	0.607 (0.037)	0.649 (0.042)	01:16:17
Sampling	KMeansSMOTE	0.679 (0.049)	0.650 (0.044)	0.657 (0.043)	0.628 (0.038)	0.654 (0.043)	00:57:02
Sampling	SMOTE-Tomek	0.685 (0.051)	0.660 (0.048)	0.648 (0.044)	0.618 (0.037)	0.653 (0.045)	00:49:41
Sampling	SMOTE-ENN	0.681 (0.062)	0.653 (0.055)	**0.665** (0.058)	***0.662** (0.054)	**0.665** (0.057)	00:27:04
Sampling	Random Oversampler	0.684 (0.050)	**0.666** (0.043)	0.645 (0.042)	0.607 (0.043)	0.651 (0.045)	01:15:33
Sampling	Random Undersampler	0.687 (0.046)	**0.676** (0.038)	0.649 (0.036)	0.632 (0.033)	0.661 (0.038)	01:00:11
Sampling	NearMiss	0.644 (0.026)	*0.581 (0.029)	*0.587 (0.024)	*0.540 (0.030)	0.588 (0.027)	00:58:41
Perturbation	Polynomial Features	0.689 (0.043)	0.654 (0.048)	0.653 (0.037)	0.593 (0.029)	0.647 (0.039)	02:00:11
Perturbation	Feature Interaction	0.686 (0.044)	0.653 (0.047)	0.654 (0.036)	*0.595 (0.028)	0.647 (0.038)	01:55:24
Perturbation	PCA	0.686 (0.047)	0.657 (0.046)	0.653 (0.038)	*0.593 (0.026)	0.647 (0.039)	01:29:40
Perturbation	Standardization	0.686 (0.045)	0.636 (0.039)	0.650 (0.041)	*0.574 (0.033)	0.637 (0.040)	01:41:28
Perturbation	Min-Max Scaling	0.686 (0.045)	0.652 (0.039)	0.651 (0.041)	0.607 (0.034)	0.649 (0.040)	01:11:39
Perturbation	Robust Scaling	0.686 (0.045)	0.661 (0.037)	0.651 (0.041)	*0.583 (0.033)	0.645 (0.039)	01:28:54
Perturbation	Log Transformation	0.686 (0.046)	0.652 (0.045)	0.651 (0.041)	0.622 (0.034)	0.653 (0.041)	01:03:10
Perturbation	Power Transformation	0.689 (0.046)	0.639 (0.047)	0.652 (0.040)	*0.569 (0.030)	0.637 (0.041)	01:42:53
Perturbation	Noise Addition	0.686 (0.048)	0.652 (0.044)	0.655 (0.041)	0.629 (0.033)	0.656 (0.042)	01:02:08
Generation	GAN	0.686 (0.048)	0.648 (0.046)	0.653 (0.039)	**0.641** (0.034)	0.657 (0.042)	02:26:18
Generation	VAE	0.682 (0.047)	0.637 (0.043)	0.652 (0.037)	0.635 (0.036)	0.652 (0.041)	12:07:27
Generation	CGAN	0.640 (0.050)	**0.665** (0.044)	0.656 (0.043)	**0.649** (0.033)	0.652 (0.042)	28:52:40
Best S + G	SMOTE-ENN + GAN	0.682 (0.060)	0.655 (0.057)	0.663 (0.056)	**0.662** (0.055)	**0.666** (0.057)	16:17:55
Best P + S	Noise Addition + SMOTE-ENN	0.684 (0.056)	0.658 (0.051)	**0.665** (0.055)	***0.667** (0.054)	**0.668** (0.054)	00:39:14
Best P + G	Noise Addition + GAN	0.687 (0.048)	0.648 (0.047)	0.655 (0.038)	**0.641** (0.036)	0.657 (0.042)	01:08:15

both *SMOTE-ENN* and *Noise Addition*, which were alone the overall best in their respective categories (+0.01 and +0.001 improvement over the baseline), did also the best together (+0.013 improvement over the baseline) among all combinations. Overall, the implied finding is that if a technique was not useful alone, then including it in a combination will not make it substantially better. However, if a technique was promising alone, then it is worth trying it in combination with another augmentation category.

4.3 Limitations

Compared to related work, a previous LA study [4] reported improved accuracy (not AUC) of SVM, RF, and neural network models when using SMOTE and ADASYN (which our results did not determine to be useful). However, that study used a smaller dataset of 480 students and did not cross-validate on an ID attribute (we used school-level cross-validation [67]), so their test set performance may have been inflated. Other prior studies that used data augmentation in LA employed different algorithms, so comparing to their results is not possible. Finally, in non-LA domains, the improvements after data augmentation ranged from substantial, to minor, to sometimes even negative [7].

Deep generation introduced more variation in AUCs across seeds and did not improve performance. These methods may need hyperparameter tuning (e.g., by random search) to be effective, and appear impractical for most LA practitioners whose focus is on model performance. Other settings might make deep generation more effective, but that would require considerable computational resources and time. Moreover, non-deterministic models combined with deep generation contain other forms of non-determinism besides the seed. We observed that during the technique chaining, even though most iterations finished fine, some could produce null values and crash on an error we did not encounter during testing.

We did not inspect feature selection because it did not change the baseline performance and substantially slowed the model fitting. Yet, feature selection could be explored for improving feature engineering *after* augmentation is applied, especially in the context of larger LA datasets.

Lastly, our results are limited to the replicated paper's context. Expanding the evaluation could reveal more general properties of augmentation, allowing the findings to be reliably generalized to another LA context.

4.4 Practical Observations and Implications for the Field of LA

Our results bear several implications for LA researchers and practitioners who seek to use data augmentation in predictive modeling. First, we observed considerable variation in how much different augmentation techniques improved – or, at times, worsened – the performance of different models. If researchers have limited time, certain techniques, such as SMOTE-ENN, might be more worth trying than others. Second, we note training runtime tradeoffs. Most performance improvements come at the cost of longer training, but, surprisingly, some techniques can both improve performance and decrease runtime. Third, often neglected in past research, it seems challenging to establish statistical significance when testing several augmentation techniques on sample sizes typical for our field. We provide open-source code (see Section 5) to facilitate rigorous significance testing for this purpose.

Our study provided new insights indicating that for binary classification, most augmentation methods are too complex to yield performance improvements without extensive hyperparameter tuning (which may be impractical given the runtimes noted here). However, sampling can slightly improve model performance, as observed with SMOTE-ENN, particularly on the non-deterministic models. Researchers can easily use this method for predictive modeling, as SMOTE-ENN is implemented in the popular Python library `scikit-learn` [50]. It is important to note that other researchers may not need to run 100 iterations; we did this only to identify the most effective methods in a statistically reliable manner. For practical applications, one iteration may suffice, converging within minutes on standard PCs.

From a technical perspective, SMOTE-ENN involves two key steps [47]: after SMOTE generates synthetic samples by interpolating between existing observations (i.e., oversampling), ENN helps clean up the dataset (i.e., regularization by noise reduction). It does this by removing instances (both original and synthetic) that are misclassified by their nearest neighbors. This step reduces noisy instances, particularly near the class boundaries, thus refining the data. Notably, SMOTE-ENN was further boosted by additionally combining it with noise addition. While, at first glance, the two methods seem to have contradictory objectives, we interpret this boost as noise addition further regularizing the dataset, boosting performance on a holdout fold during cross-validation while preventing overfitting.

This is a likely reason why this process contributed to improving the performance, given that some signal in the original dataset is noise. All the features used in the models (both cognitive and affective) only estimate real-world constructs and thus have inaccuracies, as was reported in a recent paper [40] that used a similar dataset from the same digital learning platform, ASSISTments.

Overall, our findings have implications for binary classification in LA and other prediction tasks involving SML, which are prevalent in educational data science. We also used cognitive and affective features that are common predictors in other LA studies. Therefore, researchers operating in similar contexts can try using data augmentation with SMOTE-ENN sampling, in order to increase the overall quality of training data and decrease the training time.

4.5 Open Research Challenges

Future work should evaluate the methods on other datasets and prediction tasks (beyond enrollment outcomes and binary classification). Studying other model architectures and hyperparameter settings is also warranted. We generally observed that simpler models did better, potentially because their complexity regularizes prediction more. Hence, future work may study if regularization of more complex neural networks, for example through dropout, may further boost the performance gains related to data augmentation reported here.

Another direction is to output the augmentation results, visualize how the data distribution changed after the augmentation, and examine the statistical properties of the augmented data samples. Subsequently, having such dataset allows us to directly investigate why some of the augmentation techniques work and why others do not. This includes their impact on feature importance, which we did not consider in this study.

Such research can also vary the amount of data produced by augmentation. So far, the most successful techniques in this study (sampling) produced new data samples to ensure balanced classes (i.e., up to equal class size). Our findings merit further study comparing the differences in results if, after augmentation, the size of one class exceeded the other by a certain ratio, such as 2× or 3× the size. In addition, by experimentally controlling the number of augmented data points, we can study what the most desirable ratio of synthetic to real data in LA prediction tasks is.

Researchers interested in generative models like large language models can employ them to generate additional training data, enabling learning that extends beyond the patterns present in the existing data. Related to this, future work can examine how datasets with different degrees of similarity contribute to predictive improvements. This involves measuring similarities between the original and various augmented datasets to determine how different the augmented data are and whether more different or more similar datasets improve predictions.

5 Conclusion

Using data augmentation in LA contexts is a timely topic. Much of LA research employs supervised machine learning, often using small or imbalanced data (e.g., for tasks such as dropout prediction). Within other domains, data augmentation often improved predictive performance. Yet, to the best of our knowledge, no LA paper has attempted a systematic empirical evaluation of data augmentation for outcome prediction.

Our study assessed 21 data augmentation techniques, including their 99 chained combinations. We observed that traditional sampling techniques are as good or better than more recent, advanced techniques, which take much longer to compute and are less stable due to non-determinism and more hyperparameters. Further, chaining yielded tangible, although small improvements beyond sampling alone. Lastly, the most effective technique, SMOTE-ENN, which significantly improved predictive performance, was also the fastest to run (even much faster than not performing any data augmentation). These results suggest that off-the-shelf sampling techniques may provide the most applicable improvements in LA contexts. Still, some transformation and generation techniques

worsened predictive performance, so careful method selection is necessary.

Although the overall AUC improvements may seem small, sometimes even small improvements in predictive performance matter. The practical relevance of these improvements depends on context and application (e.g., scale, risk, and benefits of correct classification) and is beyond the scope of this paper. Our contribution includes open-source materials for other researchers to conduct rigorous significance testing of improvements resulting from a suite of augmentation techniques. Ultimately, our research tested the utility of data augmentation in LA, providing insights into its applicability and the conditions under which it may or may not enhance predictive performance.

In parallel, our second contribution was independently validating the results from a previous LAK paper [67]. By independently reproducing results using our own analysis code based on the original study, we demonstrated further evidence to the LA community regarding the robustness of the study's findings. From a broader perspective, replication is a cornerstone of scientific progress. Our results serve to reinforce the earlier work, which is rare in LA research, and contribute towards addressing the reproducibility crisis [22, 23, 28].

All code used in our research and full results are available [70]. The dataset must be requested from its owners [49].

Acknowledgments

This work was supported by JST CREST Grant Number JPMJCR22D1, JSPS KAKENHI Grant Number JP22H00551, and MEXT "Innovation Platform for Society 5.0" Program Grant Number JPMXP0518071489, Japan. We thank Andres Felipe Zambrano for providing the code for the original study. This work benefited from the resources of the Institute of Advanced Computing at Tampere Center for Scientific Computing.

References

[1] ACM. 2020. Artifact Review and Badging – Current. https://www.acm.org/publications/policies/artifact-review-and-badging-current.

[2] Ma Victoria Almeda and Ryan S Baker. 2020. Predicting Student Participation in STEM Careers: The Role of Affect and Engagement during Middle School. *Journal of Educational Data Mining* 12, 2 (2020), 33–47. https://doi.org/10.5281/zenodo.4008054

[3] Ebedia Hilda Am, Indriana Hidayah, and Sri Suning Kusumawardani. 2021. A literature review of knowledge tracing for student modeling: research trends, models, datasets, and challenges. *Journal of Inf. Technology and Computer Science* 6, 2 (2021), 183–194. https://doi.org/10.25126/jitecs.202162344

[4] Usman Ashfaq, PM Booma, and Raheem Mafas. 2020. Managing Student Performance: A Predictive Analytics using Imbalanced Data. *International Journal of Recent Technology and Engineering* 8, 6 (2020), 6. https://doi.org/10.35940/ijrte.E7008.038620

[5] Ryan S. Baker. 2019. Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes. *Journal of Educational Data Mining* 11, 1 (2019), 1–17. https://doi.org/10.5281/zenodo.3554745

[6] Ryan SJd Baker et al. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68, 4 (2010), 223–241.

[7] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A Survey on Data Augmentation for Text Classification. *ACM Comput. Surv.* 55, 7, Article 146 (dec 2022), 39 pages. https://doi.org/10.1145/3544558

[8] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[9] Alan Mark Berg, Stefan T Mol, Gábor Kismihók, and Niall Sclater. 2016. The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics* 3, 1 (2016), 107–128. https://doi.org/10.18608/jla.2016.31.7

[10] Angona Biswas et al. 2023. Generative Adversarial Networks for Data Augmentation. In *Data Driven Approaches on Medical Imaging*. Springer, Cham, 159–177. https://doi.org/10.1007/978-3-031-47772-0_8

[11] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357. https://doi.org/10.1613/jair.953

[12] Kwok Tai Chui, Lap-Kei Lee, Fu Lee Wang, Simon K. S. Cheung, and Leung Pun Wong. 2024. A Review of Data Augmentation and Data Generation Using Artificial Intelligence in Education. In *Technology in Education*. Springer, Singapore, 242–253. https://doi.org/10.1007/978-981-99-8255-4_21

[13] Elizabeth B Cloude, Daryn A Dever, Debbie L Hahs-Vaughn, Andrew J Emerson, Roger Azevedo, and James Lester. 2022. Affective Dynamics and Cognition During Game-Based Learning. *Transactions on Affective Computing* 13, 4 (2022), 1705–1717. https://doi.org/10.1109/TAFFC.2022.3210755

[14] Doug Clow. 2012. The learning analytics cycle: closing the loop effectively. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)*. Association for Computing Machinery, New York, NY, USA, 134–138. https://doi.org/10.1145/2330601.2330636

[15] Belle Dang, Andy Nguyen, and Sanna Järvelä. 2024. The Unspoken Aspect of Socially Shared Regulation in Collaborative Learning: AI-Driven Learning Analytics Unveiling 'Silent Pauses'. In *Proceedings of the 14th LAK Conference*. ACM, USA, 231–240. https://doi.org/10.1145/3636555.3636874

[16] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 44, 3 (1988), 837–845. https://doi.org/10.2307/2531595

[17] Blaženka Divjak, Barbi Svetec, Damir Horvat, and Nikola Kadoić. 2023. Assessment validity and learning analytics as prerequisites for ensuring student-centred learning design. *British journal of educational technology* 54, 1 (2023), 313–334. https://doi.org/10.1111/bjet.13290

[18] Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2 (2012), 145–157.

[19] Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction* 29 (2014), 153–170. https://doi.org/10.1016/j.learninstruc.2012.05.003

[20] Brendan Flanagan, Rwitajit Majumdar, and Hiroaki Ogata. 2022. Fine Grain Synthetic Educational Data: Challenges and Limitations of Collaborative Learning Analytics. *IEEE Access* 10 (2022), 26230–26241. https://doi.org/10.1109/ACCESS.2022.3156073

[21] Dragan Gašević, Vitomir Kovanović, and Srećko Joksimović. 2017. Piecing the learning analytics puzzle: A consolidated model of a field of research and practice. *Learning: Research and Practice* 3, 1 (2017), 63–78. https://doi.org/10.1080/23735082.2017.1286142

[22] Aaron Haim, Robert Gyurcsan, Chris Baxter, Stacy T. Shaw, and Neil T. Heffernan. 2023. How to Open Science: Debugging Reproducibility within the Educational Data Mining Conference. In *16th International Conf. on EDM*. Int. EDM Soc., USA, 114–124. https://doi.org/10.5281/zenodo.8115651

[23] Aaron Haim, Stacy Shaw, and Neil Heffernan. 2023. How to Open Science: A Principle and Reproducibility Review of the Learning Analytics and Knowledge Conference. In *13th International LAK Conference*. ACM, New York, NY, USA, 156–164. https://doi.org/10.1145/3576050.3576071

[24] J A Hanley and B J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (1982), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

[25] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *International Joint Conference on Neural Networks*. IEEE, USA, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

[26] Neil T. Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *International Journal of AI in Education* 24, 4 (2014), 470–497.

[27] Genevieve Henricks, Michelle Perry, and Suma Bhat. 2024. The Relation Among Gender, Language, and Posting Type in Online Chemistry Course Discussion Forums. In *Proceedings of the 14th LAK Conference*. ACM, USA, 189–199. https://doi.org/10.1145/3636555.3636867

[28] Matthew Hutson. 2018. Artificial intelligence faces reproducibility crisis. *Science* 359, 6377 (2018), 2 pages. https://doi.org/10.1126/science.359.6377.725

[29] Stephen Hutt et al. 2024. Feedback on Feedback: Comparing Classic Natural Language Processing and Generative AI to Evaluate Peer Feedback. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM, USA, 55–65. https://doi.org/10.1145/3636555.3636850

[30] Stephen Hutt, Sanchari Das, and Ryan S Baker. 2023. The Right to Be Forgotten and Educational Data Mining: Challenges and Paths Forward. In *Proceedings of the 16th International Conf. on Educ. Data Mining*. International EDM Society,

USA, 9 pages. https://doi.org/10.5281/zenodo.8115655
[31] Nikita Jaipuria et al. 2020. Deflating Dataset Bias Using Synthetic Data Augmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, USA, 3344–3353. https://doi.org/10.1109/CVPRW50498.2020.00394
[32] Lan Jiang, Clara Belitz, and Nigel Bosch. 2024. Synthetic Dataset Generation for Fairer Unfairness Research. In *Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24)*. ACM, USA, 200–209. https://doi.org/10.1145/3636555.3636868
[33] Shiming Kai et al. 2015. A Comparison of Video-Based and Interaction-Based Affect Detectors in Physics Playground. In *Proceedings of the 8th International Conference on Educational Data Mining*. International EDM Society, USA, 8 pages. https://files.eric.ed.gov/fulltext/ED560544.pdf
[34] Shamya Karumbaiah, Ryan S Baker, Jaclyn Ocumpaugh, and Juliana Ma Alexandra L Andres. 2021. A re-analysis and synthesis of data on affect dynamics in learning. *IEEE Transactions on Affective Computing* 14, 2 (2021), 1696–1710. https://doi.org/10.1109/TAFFC.2021.3086118
[35] Tanja Käser and Giora Alexandron. 2024. Simulated Learners in Educational Technology: A Systematic Literature Review and a Turing-like Test. *International Journal of Artificial Intelligence in Education* 34, 2 (2024), 545–585. https://doi.org/10.1007/s40593-023-00337-2
[36] Ummul-Kiram Kathawalla, Priya Silverstein, and Moin Syed. 2021. Easing Into Open Science: A Guide for Graduate Students and Their Advisors. *Collabra: Psychology* 7, 1 (01 2021), 18684. https://doi.org/10.1525/collabra.18684
[37] Mark Klose, Vasvi Desai, Yang Song, and Edward Gehringer. 2020. EDM and Privacy: Ethics and Legalities of Data Collection, Usage, and Storage. In *Proceedings of the 13th International Conf. on Educational Data Mining*. International EDM Society, USA, 9 pages. https://eric.ed.gov/?id=ED607820
[38] Charles Lang, George Siemens, Alyssa Wise, and Dragan Gašević (Eds.). 2017. *Handbook of Learning Analytics* (1st ed.). Society for Learning Analytics Research (SoLAR), Alberta, Canada. https://doi.org/10.18608/hla17
[39] Qinyi Liu, Mohammad Khalil, Jelena Jovanovic, and Ronas Shakya. 2024. Scaling While Privacy Preserving: A Comprehensive Synthetic Tabular Data Generation and Evaluation in Learning Analytics. In *Proc. of the 14th LAK Conference*. ACM, USA, 620–631. https://doi.org/10.1145/3636555.3636921
[40] Xiner Liu, Ashish Gurung, Ryan S. Baker, and Amanda Barany. 2024. Understanding the Impact of Observer Effects on Student Affect. In *Advances in Quantitative Ethnography*. Springer Nature Switzerland, Cham, 79–94. https://doi.org/10.1007/978-3-031-76332-8_7
[41] Christopher J Maclellan, Erik Harpstead, Rony Patel, and Kenneth R Koedinger. 2016. The Apprentice Learner Architecture: Closing the Loop between Learning Theory and Educational Data. In *Proceedings of the 9th International Conf. on EDM*. International EDM Society, USA, 8 pages.
[42] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. 2022. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings* 3, 1 (2022), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020
[43] Josey Mathew, Ming Luo, Chee Khiang Pang, and Hian Leng Chan. 2015. Kernel-based SMOTE for SVM classification of imbalanced datasets. In *IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society*. IEEE, USA, 001127–001132. https://doi.org/10.1109/IECON.2015.7392251
[44] Michael Mogessie. 2023. alpha.correction.bh: Benjamini-Hochberg Alpha Correction. https://doi.org/10.32614/CRAN.package.alpha.correction.bh
[45] Michael Mogessie, J Elizabeth Richey, Bruce M McLaren, Juan Miguel L Andres-Bray, and Ryan S Baker. 2020. Confrustion and gaming while learning with erroneous examples in a decimals game. In *Artificial Intelligence in Education: 21st International Conference*. Springer, USA, 208–213.
[46] Alhassan Mumuni and Fuseini Mumuni. 2022. Data augmentation: A comprehensive survey of modern approaches. *Array* 16 (2022), 100258. https://doi.org/10.1016/j.array.2022.100258
[47] Mirza Muntasir Nishat et al. 2022. A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset. *Scientific Programming* 2022, 1 (2022), 3649406.
[48] Zachary A Pardos, Ryan SJD Baker, Maria OCZ San Pedro, Sujith M Gowda, and Supreeth M Gowda. 2014. Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *J. of Learning Analytics* 1, 1 (2014), 107–128.
[49] Thanaporn Patikorn, Ryan S. Baker, and Neil T. Heffernan. 2020. ASSISTments Longitudinal Data Mining Competition Special Issue: A Preface. *Journal of Educational Data Mining* 12, 2 (2020), i–xi. https://doi.org/10.5281/zenodo.4008048
[50] F. Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[51] Maria Ofelia Pedro, Ryan Baker, Alex Bowers, and Neil Heffernan. 2013. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In *Proceedings of the 6th International Conf. on Educational Data Mining*. International EDM Society, USA, 8 pages.
[52] Radek Pelánek. 2017. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction* 27, 3 (2017), 313–350. https://doi.org/10.1007/s11257-017-9193-2
[53] Vasileios C. Pezoulas et al. 2024. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal* 23 (2024), 2892–2910. https://doi.org/10.1016/j.csbj.2024.07.005
[54] J Elizabeth Richey, Juan Miguel L Andres-Bray, Michael Mogessie, Richard Scruggs, Juliana MAL Andres, Jon R Star, Ryan S Baker, and Bruce M McLaren. 2019. More confusion and frustration, better learning: The impact of erroneous examples. *Computers & Education* 139 (2019), 173–190.
[55] Cristobal Romero, Sebastian Ventura, Mykola Pechenizkiy, and Ryan SJd Baker (Eds.). 2010. *Handbook of educational data mining*. CRC Press, Boca Raton, FL, USA. https://doi.org/10.1201/b10274
[56] Maria Ofelia Z San Pedro et al. 2022. Exploring Selective College Attendance and Middle School Cognitive and Non-cognitive Factors Within Computer-Based Math Learning. In *Social and Emotional Learning and Complex Skills Assessment*. Springer, Cham, 217–247.
[57] Nabila Sghir, Amina Adadi, and Mohammed Lahmer. 2023. Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Education and information technologies* 28, 7 (2023), 8299–8333. https://doi.org/10.1007/s10639-022-11536-0
[58] Anuraganand Sharma, Prabhat Kumar Singh, and Rohitash Chandra. 2022. SMOTified-GAN for Class Imbalanced Pattern Classification Problems. *IEEE Access* 10 (2022), 30655–30665. https://doi.org/10.1109/ACCESS.2022.3158977
[59] Ruoqi Shen, Sébastien Bubeck, and Suriya Gunasekar. 2022. Data Augmentation as Feature Manipulation. In *International Conference on Machine Learning*. PMLR, USA, 19773–19808. https://proceedings.mlr.press/v162/shen22a.html
[60] Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6, 1 (2019), 60. https://doi.org/10.1186/s40537-019-0197-0
[61] Shreya Singhal, Andres Felipe Zambrano, Maciej Pankiewicz, Xiner Liu, Chelsea Porter, and Ryan S. Baker. 2024. De-Identifying Student Personally Identifying Information with GPT-4. In *17th International Conf. on EDM*. Int. EDM Society, USA, 559–565. https://doi.org/10.5281/zenodo.12729884
[62] Stefan Slater and Ryan S Baker. 2018. Degree of error in Bayesian knowledge tracing estimates from differences in sample sizes. *Behaviormetrika* 45, 2 (2018), 475–493. https://doi.org/10.1007/s41237-018-0072-x
[63] Shashank Sonkar, Xinghe Chen, Myco Le, Naiming Liu, Debshila Basu Mallick, and Richard Baraniuk. 2024. Code Soliloquies for Accurate Calculations in Large Language Models. In *Proceedings of the 14th LAK Conference*. ACM, USA, 828–835. https://doi.org/10.1145/3636555.3636889
[64] Xu Sun and Weichao Xu. 2014. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Processing Letters* 21, 11 (2014), 1389–1393. https://doi.org/10.1109/LSP.2014.2337313
[65] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Empirical Methods in Natural Language Processing*. Assoc. for Computational Linguistics, Hong Kong, China, 6382–6388. https://doi.org/10.18653/v1/D19-1670
[66] Yang Yue, Ying Li, Kexin Yi, and Zhonghai Wu. 2018. Synthetic data approach for classification and regression. In *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*. IEEE, USA, 1–8. https://doi.org/10.1109/ASAP.2018.8445094
[67] Andres Felipe Zambrano and Ryan S. Baker. 2024. Long-Term Prediction from Topic-Level Knowledge and Engagement in Mathematics Learning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM, New York, NY, USA, 66–77. https://doi.org/10.1145/3636555.3636851
[68] Chen Zhan, Oscar Blessed Deho, Xuwei Zhang, Srecko Joksimovic, and Maarten de Laat. 2023. Synthetic data generator for student data serving learning analytics: A comparative study. *Learning Letters* 1 (2023), 15 pages. https://doi.org/10.59453/KHZW9006
[69] Liang Zhang, Jionghao Lin, Conrad Borchers, Meng Cao, and Xiangen Hu. 2024. 3DG: A Framework for Using Generative AI for Handling Sparse Learner Performance Data From Intelligent Tutoring Systems. In *Companion Proceedings of the 14th LAK Conference*. ACM, USA, 12 pages.
[70] Valdemar Švábenský, Conrad Borchers, Elizabeth B. Cloude, and Atsushi Shimada. 2024. Supplementary Materials: Evaluating the Impact of Data Augmentation on Predictive Model Performance. https://doi.org/10.5281/zenodo.14257159