Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education

Lele Sha, Mladen Raković, Angel Das, Dragan Gašević, Guanliang Chen

***Abstract*—Predictive modelling is a core technique used in tackling various tasks in learning analytics research, e.g., classifying educational forum posts, predicting learning performance, and identifying at-risk students. When applying a predictive model, it is often treated as the first priority to improve its prediction accuracy as much as possible. Class balancing, which aims to adjust the unbalanced data samples of different class labels before using them as input to train a predictive model, has been widely regarded as a powerful method for boosting prediction accuracy. However, its impact on algorithmic bias remains largely unexplored, i.e., whether the use of class balancing methods would alleviate or amplify the differentiated prediction accuracy received by different groups of students (e.g., female vs. male). To fill this gap, our study selected three representative predictive tasks as the testbed, based on which we (i) applied two well-known metrics (i.e., hardness bias and distribution bias) to measure data characteristics to which algorithmic bias might be attributed; and (ii) investigated the impact of a total of 11 class balancing techniques on prediction fairness. Through extensive analysis and evaluation, we found that class balancing techniques, in general, tended to improve predictive fairness between different groups of students. Furthermore, class balancing techniques (e.g., SMOTE and ADASYN), which add samples to the minority group (i.e., over-sampling) can enhance the predictive accuracy of the minority group while not negatively affecting the majority group. Consequently, both fairness and accuracy can be improved by applying these oversampling class balancing methods. All data and code used in this study are publicly accessible via https://github.com/lsha49/FairCBT**

***Index Terms*—class balancing, machine learning, algorithmic bias**

I. Introduction

In the age of big data, predictive modelling has been applied to automate a plethora of labour-intensive tasks in education. For instance, researchers have been interested in applying Machine Learning (ML) models in online discussion forums to categorize student posts in terms of post urgency (whether it requires immediate attention from the instructor) [1], confusion (whether it requires instructor's response on a course-related question that a student asked) [2], and sentiment (whether positive or negative emotions emerge from discussion posts) [3]. Researchers have also applied ML models to preempt student dropout by identifying students who are likely to fail a course based on their interactions with different course resources in digital learning environments, e.g., attempting weekly quizzes and viewing lecture videos [4]. Other applications include automated assessment grading [5], identification of discourse and argumentative elements in student writings [6], and knowledge tracing in students' learning processes over semester [7]. Based on this increasing trend, we can only expect interest in predictive models, particularly those dealing with classification problems, to continue growing in the learning analytics research community in the coming years.

Often, researchers need to address the problem of class imbalance in real-world datasets, where there is an unequal ratio of different classes in the data samples - e.g., when detecting cognitive presence applied to online discussion messages [8], [9], over 30% of the messages are of the exploration type cognitive label, and only 6% are of the resolution type label. Similarly, when predicting students who are unlikely to complete a MOOC, the number of students who successfully completing the course hardly ever exceed 25% [10]. One promising approach for mitigating this problem is to apply Class Balancing Techniques (CBTs) to over-sample the class labels with less data samples or under-sample the class labels with more data samples to help reach parity in the training data [11]. CBTs have traditionally been applied to help datasets reach comparable amount in different classes [12], [13]. As a result, higher prediction accuracy of the minority class samples can be attained by training models with re-sampled datasets [14], [15].

While researchers continue to achieve new accuracy benchmark across different educational modelling tasks including forum post classifications [16] and dropout predictions [4], the impact of such models' predictions on students in online education and learning experience has raised concerns about the biases of the ML models [17], [18]. At times, unintentional discrimination (i.e., biased predictive accuracy against certain groups) may be happening due to factors like demographic imbalance, where certain demographic groups (e.g., gender, race) are under-represented in the dataset [19]. Gender imbalance (i.e., the data samples pertinent to students of different genders are unequal), for instance, is common across various courses and learning platforms [20], [17]. Previous research has shown that predictive models that were trained using datasets with a small sample size on distinct groups of people, such as those of different genders, race, or language groups, can cause and

L. Sha was with the Centre for Learning Analytics, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, e-mail: lele.sha1@monash.edu.

M. Raković was with the Centre for Learning Analytics, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, e-mail: mladen.rakovic@monash.edu.

A. Das was with Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, e-mail: adas0008@student.monash.edu.

D. Gašević was with the Centre for Learning Analytics, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, e-mail: dragan.gasevic@monash.edu.

G. Chen was with the Centre for Learning Analytics, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, e-mail: guanliang.chen@monash.edu. *(Corresponding author: Guanliang Chen)*

reinforce unintended bias [17], [19]. Unfortunately, samples from different groups rarely reach comparable quantity in most real-world data collections in education [20], [21], e.g., female students have been traditionally under-represented in STEM courses with only around 1 in 5 learners in a STEM MOOC being female [21]. Given that the predictive models have been widely adopted to empower teaching effectiveness — e.g., instructors may offer a timely intervention to students based on the results of dropout prediction and instructors may also provide forum responses to those classified as urgent posts — biased prediction resulted from under-represented demographic groups in training data may cause unintentional bias in teaching support.

In recent years, researchers proposed that using CBTs could also potentially mitigate the issue of under-represented groups in datasets and enhance fairness of predictive models [19], [8], [22], [23]. However, to date, the effectiveness of such approaches still remains relatively unexplored, due to the lacuna in research that aims to systematically evaluate effectiveness of existing CBTs of different types (e.g., under-sampling and over-sampling) when applied to representative educational modelling tasks.

Hence, in this study we sought to assess the impact of popular state-of-the-art CBTs on algorithmic bias in predictive modelling tasks in education. Formally, this study was guided by the following **R**esearch **Q**uestion:

RQ To what extent can class balancing techniques affect fairness by re-balancing datasets to minimise the predictive parity between different demographic groups?

To answer the RQ, we conducted experiments on three long-standing educational modelling tasks, namely student dropout prediction, student performance modelling, and forum post classification. We first evaluated to what extent bias existed in original datasets in terms of demographic distributions of different groups (denoted as *distribution bias*) and difficulty levels of making a correct prediction across different groups (denoted as *hardness bias*). Then, for each task, we replicated the model that has been well-documented in the literature and has been shown to achieve the benchmark prediction accuracy in previous studies. We selected a Convolutional Neural Network (CNN) based model [4] for the student dropout prediction, a Random Forest model for the student performance modelling, and an CLSTM (Convolutional Long Short-Term Memory) model [24] for the forum classification task. We selected a total of 11 representative CBTs of under- and/or over-sampling and input re-balanced samples to train the selected models. This study was approved by the Human Research Ethics Committee at an Monash university (Project ID 30074). The main findings are as follows:

- Bias in terms of distribution and hardness generally existed in the three educational datasets prior to model training. Furthermore, hardness bias tended to be associated with algorithmic unfairness.
- Most of the CBTs tended to improve fairness of the prediction results in at least two out of three educational modelling tasks. In particular, three CBTs — i.e., Tomek-Link, SMOTE and SMOTE-TomekLink — improved fairness in all tasks.
- When fairness was improved by adding more samples to the under-represented group, the overall prediction accuracy of the dataset also tended to improve.

II. Related work

A. Predictive Modelling in Learning Analytics

As a result of the development and use of various online tools and systems for distance learning, rich student activity data are generated. By applying machine learning based models, researchers and practitioners are able to solve various challenges in learning analytics that were traditionally difficult to tackle. Joksimović et al. [25] categorised two core research areas in learning analytics, namely predictive analytics and discourse analytics. The aim of predictive analytics is to facilitate student learning by predicting their learning outcomes in advance. The most prevalent use case has been to mitigate the notoriously criticised low completion ratio of MOOCs [26], [4]. Researchers have devoted their attention to predicting student performance to preempt their risks of dropout [4], [26], i.e., students at risk of failing or dropping out of a course are identified based on data about their learning activities. Consequently, instructors may intervene and provide necessary support to those students. Discourse analytics are mainly concerned with the analysis of textual data generated by students and instructors to facilitate teaching and learning. A notable example of discourse analytics is automatic classification of online forum posts. While the adoption of online discussion forums can help form a learning community that supports effective student-student and student-instructor interaction much like a traditional classroom-based learning environment [27], the sheer volume of posts to online discussion forums generated by tens of thousands of students may overwhelm teaching teams. The research on automatic classification of forum posts produced computational models that can effectively identify relevant posts for instructors, and enable them to provide a timely and tailored response to students. Previous studies proposed applying automatic classification models for analyzing various online discussion messages such as to identify course content-related questions [28], to detect confusion in students' posts [29], to detect student achievement emotions [30], to detect social and cognitive presence [31], and to identify the level of urgency expressed by students [1].

Despite rigorous efforts by various researchers in predictive modelling in education, existing studies often stressed accuracy as the main dimension to improve. Take forum post classification for example. Traditional ML models such as Support Vector Machine (SVM) and Gradient Boost (GB) have been widely applied and a variety of studies proposed novel features to be engineered (e.g., hash tags on discussions [32], domain specific words [33]) to attain improved accuracy. In recent years, researchers have moved forward to exploit developments of more powerful deep learning (DL) models (e.g., LSTM) and pre-trained language models (e.g., BERT) to seek further accuracy improvement for various forum classification tasks like urgency [16], sentiment [34] and question-identification [24].

B. Dealing with imbalanced datasets

One common problem that predictive modelling face is imbalanced datasets, which can cause under-represented minority class samples to be insufficiently trained by the ML model, hence undermining the predictive accuracy of the model [35], [36]. Previous studies [15], [37], [14] have proposed to balance the majority and minority class samples by applying CBTs to re-sample the original dataset so as to boost the predictive accuracy. For example, Cavalcanti et al. [14] applied SMOTE to re-balance samples of different feedback levels and used the balanced samples to predict whether the feedback provided by instructors was of a high quality. The authors achieved an accuracy of up to 87%, which had around 2% improvement over the model without SMOTE. In general, researchers typically apply three strategies to help reach parity among different class samples, namely, under-sampling the majority group, over-sampling the minority group, and a hybrid of under- and/or over-sampling. A summary of the popular CBTs is provided in Table I.

The easiest under-sampling CBT is to remove samples from the majority group randomly, denoted as random under-sampling [38]. However, researchers have noted that down-sizing the majority group randomly could easily lead to the dataset losing meaningful information which then can cause predictive accuracy to decline [11], [38]. More sophisticated approaches have been designated to avoid this shortfall. For instance, Condensed Nearest Neighbors [39] is among one of the earliest CBTs which iteratively decides whether a sample should be deleted based on one nearest neighbour. Further, a CBT named Edited Nearest Neighbors [40] utilised k nearest neighbors which can help identify noisy or borderline samples and remove them. Another approach, Near Miss [41] selects samples to remove by average distances of the nearest neighbours. Lastly, TomekLink [42] identifies pairs of nearest neighbors and removes the majority sample until there is no minimally distanced nearest neighbor pairs of major and minor group.

In contrast to under-sampling, over-sampling CBTs aim to add samples to the minority group to help reach a comparable sample size with the majority group. Synthetic Minority Over-sampling Technique or SMOTE is probably the most popular over-sampling CBT which is usually regarded as one of the most powerful approaches for over-sampling [12]. However, researchers noted that SMOTE-generated samples may be subject to noise issues [35], [43]. To remedy this, a few SMOTE variants have been proposed. For example, SMOTE (K-means) [35] uses k homogeneous subgroups to eliminate generated noise, and borderline SMOTE [43] only generates minority samples at borderline to prevent noisy samples. Lastly, as an alternative to SMOTE-based approaches, ADASYN [44] is proposed to generate more samples based on the different neighborhood rule which is shown to be more effective than SMOTE in terms of noise reduction.

Either over-sampling or under-sampling approaches have been shown to be effective when used alone [12], [37], [14], [15]. However, in recent years, a combination of multiple under- and/or over-sampling CBTs have been proven to achieve better accuracy performance for imbalanced datasets [36], [45]. Existing studies have proposed combinations of several popular CBTs such as one over-sampling and one under-sampling (e.g., SMOTE and TomekLink) [45], the combination of Tomek's link with SMOTE has been shown to be able to eliminate excessive noise from SMOTE [46]. Another popular CBT combination is to apply two under-sampling (e.g.,Tomek's link and Condensed Nearest neighbor) which is formally known as One-Sided Selection [47], and the resulting sampling process can have the advantages of both technique. For example, the ambiguous link points are removed by Tomek's link, and then the redundant points from decision boundary are removed by Condensed Nearest Neighbor.

C. Data imbalance and bias

Even prior to model training, bias may exist in datasets. Existing research notes that one important factor causing prediction bias is data imbalance where certain demographic group (e.g., male students) is over or under-represented in the datasets [17], [19]. Given the impact of dataset bias on predictive fairness, Yan et al. [19] suggested two important metrics to quantify bias in a dataset, namely distribution and hardness bias. Specifically, distribution bias refers to a scenario where there is an uneven distribution of samples from different demographic groups, e.g., there is more male samples than female samples. This may cause the model to be insufficiently trained on the minority group samples, which, in turn, may result in unsatisfactory prediction performance relative to the minority group. However, even with an equal sample distribution among different groups (e.g., the same number of male and female samples), the dataset may still be biased in terms of hardness, i.e., samples from one group may be more difficult to predict correctly than those from the other group. Yan et al. [19] propose to use k-Disagreeing Neighbors (kDN) to measure the hardness bias. To put it simply, kDN measures to what extent an instance overlap (in terms of Euclidean distance) with its k nearest neighbors which do not belong to the same class.

When it comes to the field of learning analytics, existing works on measuring predictive bias are typically based on one or several of the following definitions of algorithmic fairness including *statistical parity* [52], *equalised odds* [53], and *equalised opportunities* [53]. For instance, when trying to detect bias in modeling student performance, most of the existing studies have focused on measuring whether there is a predictive disparity between different demographic groups in line with the definition of *equalised odds/opportunities* [17], [54], [55], [18], [56]. To the best of our knowledge, only one study [57] systematically evaluated data bias (i.e., an unequal representation of different groups of people/students in the training data) alongside predictive bias in an educational context. In their work, Deho et al. [57] debiased training data during the data pre-processing step and ensured that the numbers of samples/records specific to native English and non-native English students are equal in the training data, which are then used to train models for student dropout prediction. The results show that both the baseline model and the fairness-aware models tend to replicate data biases in the prediction

TABLE I: A summary of the class balancing techniques. The techniques included in this review are denoted as *Yes* under the column *Included*.

Category	Techniques	Description	Used in studies	Included
Under sampling	Tomek's links	Remove samples identified on the Tomek's link.	[42], [23]	Yes
	Edited Nearest Neighbours (Edited-NN)	Remove samples by nearest neighbours rule.	[40], [23]	
	NearMiss	Remove samples based on the average distance from minority class samples.	[41], [23]	
	Condensed Nearest Neighbour (Condensed-NN)	Remove samples far away from the decision neighbor iteratively.	[39]	
	Random under-sampling	Randomly remove majority samples.	[19]	No
	Clustering-based under-sampling	Remove samples by their representative clustering group.	[48]	
	Inverse random under-sampling	Severely under sample the majority class based on clustering group.	[49]	
	Repeated Edited Nearest Neighbours	Mutli-class resampling method by Nearest Neighbours rule.	[50]	
Over sampling	SMOTE	Create synthetic samples of minority class from their k nearest neighbours.	[12], [19], [14], [23], [9]	Yes
	SMOTE (K-means)	Apply K-means clustering before SMOTE.	[19], [35]	
	SMOTE (Borderline)	Add samples from borderline data points.	[43]	
	ADASYN	Add samples adaptively by distribution.	[44]	
	Random over-sampling	Randomly add minority sample.	[19]	No
	ML SMOTE	A multi-label SMOTE variant.	[51]	
Hybrid	SMOTE-TomekLink	Combine SMOTE and TomekLink.	[23]	Yes
	One-Sided Selection	Combine TomekLink and the Condensed Nearest Neighbor.	[47]	
	Pipeline	Pipeline of multiple samplers.	[23]	

results. However, Deho et al. [57] focused mostly on the distribution side without trying to tackle other aspects of data bias (e.g., hardness bias). Consequently, the association between bias at source (i.e., data bias) and predictive outcome (i.e., model predictions) may not be fully revealed. We note similar trend in the broader ML community [58], [59]. In contrast to existing studies, we conducted a study that included an evaluation of dataset characteristics in terms of distribution and hardness bias in three longstanding predictive modelling tasks in learning analytics. We then further investigated the impact of CBTs on these dataset characteristics and the subsequent algorithmic fairness/bias.

III. PRELIMINARIES

A. Educational Predictive Tasks, Datasets, and Models

In this study, we have adopted three datasets, one dataset for an educational predictive modelling task. All three tasks were binary predictions (i.e., the prediction label was either true or false), two of which were from predictive analytics (i.e., dropout prediction and performance prediction) and one task was from discourse analytics (i.e., classification of online forum discussions). For all datasets, we focused on evaluating the fairness of predictive performance between different gender groups (i.e., male and female students), given the prevalence of gender imbalance in higher education and MOOCs as noted in Section II-B. In terms of models, predictive modelling tasks typically adopt either traditional machine learning (ML) (e.g., Support vector machine, Random Forest) or deep learning (DL) (e.g., Long Short-Term Memory, Convolutional Neural Network) models. We selected a representative model per predictive modelling task by considering whether: (i) the model is used and thoroughly documented in the previous work; and (ii) the model has been shown to achieve the benchmark prediction performance. The details of each task, datasets and models are provided below.

Forum post classification: Moodle dataset. The Moodle dataset included 3,703 forum discussion posts authored by Monash University students in subjects including arts, business, design, computer science, economics, and engineering.

TABLE II: The descriptive statistics of the dataset used in this study. The columns **Male**, **Female**, show the number of forum posts generated by male and female students, respectively.

Datasets	Modelling tasks	Attributes	All	Male	Female
Moodle	Forum post classification	# Posts	3,703	1,478	2,225
		# Words	485,737	171,768	308,087
		# Avg. words / post	131.39	116.77	138.90
		# Unique words	268,824	97,004	170,171
		# Avg. unique words / post	72.71	65.94	76.72
KDDCUP (2015)	Student dropout prediction	# Enrollments	180,785	124,742	56,043
		# Dropout	148,118	105,164	42,954
		# Video activities	1,208,821	821,998	386,823
		# Avg.video activities	6.69	6.59	6.90
		# Web page activities	6,878,176	4,539,596	2,338,580
		# Avg.web page activities	38.05	36.39	41.73
OULA	Student performance modelling	# Enrollments	24,806	13,544	11,262
		# Pass	12,362	6,619	5,743
		# Clicks	73,301	42,739	30,562
		# Avg.clicks	2.95	3.16	2.71

To ensure the labels' correctness, we first engaged a junior teaching staff member to manually label the posts, and then hired two senior teaching staff to independently evaluate and correct the labels to reach high agreement of annotated labels. Using the Moodle dataset enabled us to perform a classification task of identifying whether a post is relevant (e.g., "What is the difference between non-complex and complex number?") or irrelevant (e.g., "How do I access lecture zoom link?") relative to the course content. Therefore, the classes contained in the labels are binary, either true for content-relevant, or false for content-irrelevant.

To perform forum post classification, we replicated a popular CLSTM model inspired by previous studies [60], [61], [24], [16], which reported state-of-art prediction accuracy of 80% to 90% depending on datasets used. Since DL models require at least tens of thousands of post data to be sufficiently trained, Moodle dataset alone cannot suffice the training of this model. Therefore, we applied a pre-trained language model called BERT [62] to generate post embeddings which is a well-investigated approach in previous research and the broader natural language processing (NLP) community [34], [63], [64]. The existing research has shown that BERT a) achieved superior performance over other embedding generation methods and b) can be coupled with a task model to fine-tune the model hyper-parameters. This co-training strategy is shown to be able to tune hyper-parameter with only a few thousand data samples [62], [64], [16], which was applicable to our Moodle dataset.

Student dropout prediction: KDDCUP (2015) dataset. The KDDCUP dataset is made up of 39 courses which enrolled 180,785 students in the popular online learning platform XueTangX. The dataset was used in the KDDCUP competition in 2015 which attracted in total 821 participants to predict student dropout using the dataset. The benefit of including KDDCUP dataset is twofold. First, this is one of the few public MOOC datasets that published demographic information of students (e.g., gender), and therefore, enabled fairness evaluation of predictive performance among different demographic groups. Second, KDDCUP is a widely used public datasets with extensive modelling efforts, and hence, the dataset allowed a replication of an existing model which can bring deeper understanding of the previous work [4]. The classes contained in the labels are binary, either true which indicates that student dropped out of the course, or false which means non-dropout.

To our knowledge, [4] proposed the state-of-the-art model for this dataset, i.e., the Context-aware Feature Interaction Network (CFIN) model, which proposed a context-smoothing structure to handle feature augmentation, embedding, and feature fusion, followed by an attention-based interaction adopting an CNN model to learn and predict dropout probability. The CFIN model demonstrated around 90% AUC performance, outperforming other models by around 2% in dropout prediction using the KDDCUP dataset. Therefore, we replicated this work with the same feature inputs engineered from student activity data including: video, forum and assignment. The model is available in the public repository[1].

Student performance modeling: OULA dataset. The OULA (Open University Learning Analytics) dataset included information from courses taught at the Open University (OU)

[1]https://github.com/wzfhaha/dropout_prediction

[65]. Similar to KDDCUP, the OULA dataset was one of the few public datasets which published student demographic information. The dataset included aggregated clickstream data from 24,806 students' interactions in the Virtual Learning Environment (VLE) on 22 courses. The classes contained in the labels are binary, either true which indicates student passed the course, or false which indicates student did not pass the course.

Given that OULA is a relatively recent dataset compared to KDDCUP, fewer studies have been published using the OULA dataset. One of which is open university analytics, which used the Random Forest model and reported around 80% prediction accuracy in student performance modeling with engineered features from a range of attributes such as module click stream activity, student age band and number of previous attempts, etc. The work has a public repository[2].

We summarized the three datasets statistics in Table II.

B. Evaluation Metrics

Data bias. We used two metrics to measure potential bias hidden behind the training data, including:

- **Distribution bias**, which refers to the distribution difference between samples from different demographic groups, e.g., the different numbers of training samples pertinent to male and female groups in a dataset. Therefore, we evaluated distribution bias by measuring the distribution of male and female samples of original datasets, and compared the results with that of the balanced (via CBTs) counterpart.
- **Hardness bias**, which refers to the degree to which data instances contained in a dataset are difficult to be correctly labelled. More specifically, if a data instance does not share the same task label with most of its *k-nearest neighbors*, then it tends to be difficult to correctly label this instance. Similar to the work presented in [66], [19], we used *k-Disagreeing Neighbors* (kDN) to measure the local overlap of a data instance with its k-nearest neighbors (identified by calculating their Euclidean distance) regarding their task labels. A large kDN (close to 1) indicates that the data instance is difficult to be correctly classified. We chose $k = 5$ to calculate the kDN of an instance (as suggested in [19]). If the kDN distribution of a student group (e.g., female) is different from that of the other group (e.g., male), there exists hardness bias between the two groups, which can be calculated by applying the Jensen-Shannon distance (a symmetric version of the Kullback–Leibler divergence with a finite value).

Predictive accuracy. Similar to previous studies on the three tasks we investigated [4], [17], [67], we also adopted Area Under the Curve (AUC) to measure the predictive accuracy of a ML model.

Predictive fairness. We measured predictive fairness by using the metric Absolute Between-ROC Area (ABROCA) [17], which has been widely adopted in previous studies [68], [69], [70]. Similar to other group-level fairness metrics (e.g., equalised odds/opportunity [53]), ABROCA measures the difference of the predictive accuracy between groups of students (e.g., male and female) as the bias displayed by an ML model. ABROCA is calculated by computing the definite integral between the ROC curves of two protected groups, representing the absolute difference between the two ROC curves at all thresholds. Therefore, the entire range of thresholds is accounted for rather than just using a single fixed threshold. This enables the calculation of ABROCA to be independent of the modelling process and the ABROCA value can be easily computed as a post-prediction step after ROC calculation. Note that since ABROCA represents the total ROC area difference between two group (e.g., male and female groups), a larger ABROCA value is indicative of a higher predictive bias.

[2]https://github.com/gogoladzetedo/Open_University_Analytics

IV. Methods

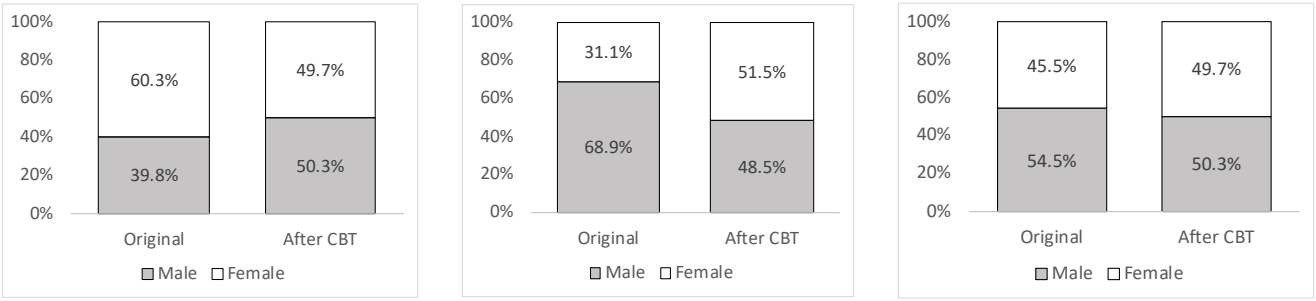
A. Class balancing techniques

Existing CBTs can be categorized into three groups: under-sampling, over-sampling and a hybrid of under- and/or over-sampling, as summarized in Table I. As we aimed to enable an in-depth understanding of the impact of class balancing techniques on algorithmic fairness, we chose a set of representative techniques from each group based on the following criteria: (i) the techniques are widely adopted and replicated by different researchers; (ii) as we tackled binary classification tasks in this study (i.e., content-relevant vs. content-irrelevant), we only included single-class sampling techniques (as opposed to multi-class); and (iii) random sampling techniques (e.g., random under- and over-samplings) were excluded given that they could potentially discard useful data and cause overfitting as noted in [38], [36]. Based on the selection criteria, we included (i) four under-sampling techniques: *Tomek's links*, *Near Miss*, *Edited Nearest Neighbour*(Edited-NN) and *Condensed Nearest Neighbour*(Condensed-NN); (ii) four over-sampling techniques: *SMOTE*, *SMOTE (K-means)*, *SMOTE (Borderline)* and *ADASYN*; (iii) three combined sampling techniques: *SMOTE Tomek's links* (one over-sampling and one under-sampling where the dataset first went through the over-sampling process, then noisy/redundant samples were removed by applying an additional under-sampling step), and *One-Sided Selection* (combined two under-sampling techniques, where the ambiguous link points were first removed by *Tomek's link*, then the redundant points from decision boundary removed by *Condensed Nearest Neighbor*); and (iv) a combination of the top-4 best performing CBTs into a single pipeline (i.e., combined two best under-sampling and two best over-sampling into a pipeline of re-sampling processes, where over-sampling CBTs were applied first, followed by the under-sampling CBTs to eliminate noise from the generated samples).

B. Study Setup

Data pre-processing involved the following steps:

- For all our experiments, we first extracted model input (i.e., features for the KDDCUP and OULA datasets, and embeddings for the Moodle dataset) for each predictive modelling task.



(a) Moodle original and after-CBT male vs. female distribution (b) KDDCUP original and after-CBT male vs. female distribution (c) OULA original and after-CBT male vs. female distribution

Fig. 1: Distribution of original samples and class balanced samples

- For KDD dropout prediction task, a total of 22 features were extracted from student activity data including: number of views of video content, questions and answers posted on a forum and average correct answers and number of revises in an assignment.
- For the OULA student performance modeling task, a total of 112 features were extracted from a range of attributes such as module click stream activity, student age band, and the number of previous attempts.
- For the Moodle forum post classification task, we first pre-processed the text contained in a post by 1) eliminating invalid characters; 2) removing stopwords; and 3) using the Natural Language Toolkit to apply word stemming. Then, BERT embedding of posts were generated by using `Bert-as-service`[3].
- After feature inputs were extracted, we divided them into the training/testing split with 80%/20% ratio, with 80% being the training set while 20% being the testing set. During training, 10% of the training data was chosen at random as validation data, and the best model was chosen based on the validation error.

Class balancing. We performed class balancing on the training set to ensure parity between the demographic (i.e., gender) groups. The class balancing techniques used are detailed in Table I as included. In line with previous studies [19], [23], [14], all CBTs were implemented using popular Python library imbalanced-learn[4].

Model implementation. For KDDCUP dropout prediction, we replicated the CFIN model from [4] repository[5]. The model contained a Deep Neural Network (DNN) based attention layer and a CNN based context-smoothing layer. The DNN was implemented with 32*32 deep layers with dropout ratio set to 0.5. The CNN was implemented with 32 convolution filters with a width of 8. During training, (i) the learning rate was set to 0.001 with Adam optimizer; (ii) epoch was set to 10 and batch size set to 256; (iii) loss type was set to logloss. At the conclusion of each epoch, 10% of the training data was chosen at random to serve as validation data. Based on validation error, we choose the best model.

For OULA student performance modeling, we replicated the Random Forest model from Open University Analytics[6]. The model was implemented using the sklearn package[7]. We applied the GridSearchCV package to find the best model hyperparameters to optimise the performance.

For the Moodle forum post classification task, the CLSTM model was implemented using *tensorflow* [8]. The model parameters were implemented in line with previous similar studies in [60], [61], [24], [16]. The dimension of input layer was set to 768 (BERT embedding input size) with 1 hidden units in the final sigmoid output layer, with L2 regularizer lambda set to 0.001. We used 128 convolution filters with a width of 5 for the CNN network, and 128 hidden states and 128 cell states for the LSTM network. During training: (i) the batch size was set to 32, and the maximum input text was set to 512; (ii) the one cycle policy was used for training, with the maximum learning rate set to 2e-05; (iii) the dropout probability was set to 0.5; and (vi) the maximum training epochs were set to 50, with shuffling at the end of each epoch and an early stopping mechanism. At the conclusion of each epoch, 10% of the training data was chosen at random to serve as validation data. Based on validation error, we choose the best model.

V. Results

We first report the extent of which bias existed in the original dataset. Then, we report the findings of under-sampling, over-sampling and hybrid CBTs to re-balance the datasets to evaluate the predictive parity between male and female groups. The results are detailed in Table III in terms of AUC (measuring accuracy) and ABROCA (measuring fairness). We also presented in Table IV a summary of whether CBTs contributed to fairness and/or accuracy improvements.

A. Data bias

In line with previous studies [19], we used two bias measurements to evaluate bias in the original datasets prior to model training, namely distribution and hardness bias. Recall distribution bias refers to a scenario where the demographic groups differ in sample size and can cause minority group

[3]https://github.com/hanxiao/bert-as-service
[4]https://imbalanced-learn.org/
[5]https://github.com/wzfhaha/dropout_prediction
[6]https://github.com/gogoladzetedo/Open_University_Analytics
[7]https://scikit-learn.org/
[8]https://www.tensorflow.org/

TABLE III: Results on Hardness bias (denoted as H-bias), AUC score and ABROCA. The top 3 best results are in bold. The fraction in the bracket indicated percentage increase/decrease compared to the baseline. The signs ↑ and ↓ are used to indicate whether a higher (↑) or lower (↓) value is more preferred in a metric.

Categroy	CBTs	Moodle			KDDCUP (2015)			OULA		
		↓ H-bias	↑ AUC	↓ ABROCA	↓ H-bias	↑ AUC	↓ ABROCA	↓ H-bias	↑ AUC	↓ ABROCA
Original	Baseline	0.021	0.842	0.068	0.034	0.882	0.042	0.021	0.799	0.041
Under sampling	NearMiss	0.025 (19.05%)	0.835 (-0.65%)	0.074 (8.82%)	0.047 (38.24%)	0.880 (-0.25%)	0.028 (-32.38%)	**0.013** (-38.10%)	0.794 (-0.70%)	**0.023** (-44.12%)
	Edited-NN	0.030 (42.86%)	0.831 (-1.04%)	0.084 (22.94%)	**0.015** (-55.88%)	0.877 (-0.61%)	**0.019** (-55.24%)	0.027 (28.57%)	0.785 (-1.78%)	0.043 (4.90%)
	Condensed-NN	0.027 (28.57%)	0.824 (-1.88%)	0.075 (10.00%)	0.030 (-11.76%)	0.882 (-0.07%)	0.032 (-23.81%)	0.020 (-4.76%)	0.794 (-0.68%)	0.029 (-28.43%)
	TomekLink	0.026 (23.81%)	0.831 (-1.05%)	**0.060** (-12.35%)	0.030 (-11.76%)	0.879 (-0.35%)	**0.026** (-37.14%)	0.023 (9.52%)	**0.799** (-0.04%)	**0.024** (-41.18%)
Over sampling	SMOTE	0.015 (-28.57%)	0.833 (-0.79%)	0.069 (1.18%)	0.018 (-47.06%)	**0.890** (0.85%)	0.028 (-34.29%)	0.015 (-28.57)	**0.800** (0.09%)	0.040 (-1.96%)
	SMOTE (K-means)	**0.008** (-61.90%)	**0.843** (0.31%)	**0.058** (-14.71%)	0.0029 (-14.71%)	**0.887** (0.53%)	0.049 (17.14%)	0.027 (28.57%)	0.798 (-0.23%)	0.050 (21.57%)
	SMOTE (Borderline)	0.017 (-19.05%)	0.834 (-0.74%)	0.080 (18.24%)	0.022 (-35.29%)	0.885 (0.35%)	0.034 (-19.05%)	0.014 (-33.33%)	0.798 (-0.23%)	0.036 (-10.78%)
	ADASYN	**0.001** (-95.24%)	**0.844** (0.44%)	**0.055** (-19.41%)	0.033 (-2.94%)	0.883 (0.08%)	0.048 (14.29%)	0.015 (28.57%)	0.797 (-0.25%)	0.030 (-26.47%)
Hybrid	SMOTE-TomekLink	**0.014** (-33.33%)	**0.840** (0.05%)	0.064 (-6.47%)	**0.014** (-58.82%)	**0.887** (0.56%)	**0.026** (-37.14%)	**0.011** (-47.62%)	0.795 (-0.59%)	0.029 (-28.43%)
	One-Sided Selection	0.028 (33.33%)	0.822 (-2.17%)	0.074 (8.24.%)	0.032 (-5.88%)	0.881 (-0.10%)	0.042 (-0.95%)	**0.010** (-52.38%)	0.798 (-0.19%)	**0.018** (-54.90%)
	Pipeline (Top-4 CBTs)	0.022 (4.76%)	0.839 (-0.12%)	0.081 (19.41%)	**0.015** (-55.88%)	0.886 (0.41%)	0.032 (-23.81%)	0.022 (4.76%)	**0.802** (0.26%)	0.049 (19.61%)

(which has less samples than majority group) to be under-represented in the dataset. Hardness bias is a kDN-based measure which quantifies how difficult it is to predict an instance correctly. We summarised our results of distribution bias of original and after-CBT in Figure 1, due to the limited space for this paper, we only include the results of applying the best CBT (i.e., ADASYN, Edited-NN and One-Sided Selection for Moodle, KDDCUP and OULA respectively) based on the fairness performance, however we noted similar results were observed by other CBTs for distribution bias. The hardness bias of all CBTs were summarised in Table III denoted as H-bias.

The prevalence of data bias. Both Moodle and KDDCUP showed distribution bias in the original dataset samples, where over 60% of forum posts in the Moodle dataset were authored by female students (Figure 1 (a) Original), and over 68% in KDDCUP were male student data (Figure 1 (b) Original). Overall, all three datasets contained distribution and hardness bias to some degree, OULA had the lowest distribution bias with only 9% gap in male and female samples sizes, while Moodle and OULA had the lowest hardness bias of 0.021.

The power of data balancing. We then applied CBTs to the three datasets. Overall, all three datasets demonstrated improvements over distribution bias (based on the *after CBT* bar chart in Figure 1), after applying CBT, male and female samples had almost the same distribution. For the hardness bias, all three datasets showed improvement ranging from 30% to 90% in the top-3 most improved H-bias (denoted as bold) compared to the original samples shown in Table III. We note that OULA showed the least change both in terms of distribution bias and hardness bias (in top-3 best H-bias) after the application of CBT. This may be due to the fact that OULA had a smaller distribution bias originally in comparison to the other two datasets, and therefore less samples were removed/added when applying CBTs and resulted in a smaller change compared to the other two datasets.

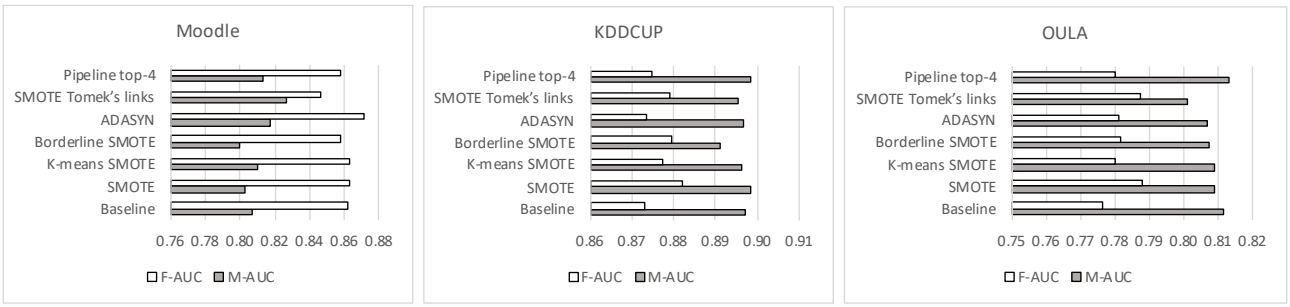
TABLE IV: Results on improvement of fairness and accuracy.

Category	CBTs	Improved Fairness			Improved Accuracy		
		Moodle	KDDCUP	OULA	Moodle	KDDCUP	OULA
Under sampling	NearMiss		√	√			
	Edited-NN		√				
	Condensed-NN		√	√			
	TomekLink	√	√	√			
Over sampling	SMOTE		√	√		√	√
	SMOTE (K-means)	√			√	√	
	SMOTE (Borderline)		√	√		√	
	ADASYN	√		√	√	√	
Hybrid	SMOTE-TomekLink	√	√	√	√	√	
	One-Sided Selection		√	√			
	Pipeline (Top-4 CBTs)		√			√	√

B. Predictive fairness and accuracy

Given the promising findings in the distribution and hardness bias, we were motivated to use the samples changed with CBTs to train respective models in each of the three predictive modelings task and present the prediction results in terms of both accuracy (measured by AUC) and fairness (measure by ABROCA). The results of AUC and ABROCA of the three predictive tasks are reported in Table III. We also summarised the results in terms of whether each CBT showed fairness improvement and/or accuracy improvement in Table IV. Note that the accuracy improvement was represented by an increase over the baseline AUC after the application of CBTs, while fairness improvement was represented by a decrease over the baseline ABROCA after application of CBTs (since ABROCA was a total difference of ROC areas between male and female, a smaller ABROCA indicated more parity between the male and female groups in terms of predictive accuracy).

The impact of data balancing on predictive fairness. We observed that most of the CBTs helped improve fairness in educational modelling tasks (i.e, 8 out of the 11 CBTs improved fairness in at least two out of the three tasks). In particular, TomekLink and SMOTE-TomekLink were shown to improve fairness across all three educational tasks (i.e., Moodle, KDDCUP and OULA). Further, the best performing CBTs in KDDCUP and OULA lead to over 50% of the



(a) (b) (c)

Fig. 2: AUC score of male and female in over-sampling CBTs

ABROCA reduction, which indicates that the gap in predictive parity between male and female students was reduced by more than 50%. Compared to KDDCUP and OULA, the best performing CBT in Moodle had smaller improvement of around 20% of the reduction in ABROCA. One reason may be that the Moodle dataset had only 3,703 posts included, significantly smaller in size compared to KDDCUP and OULA (see Table II). Therefore, less samples were generated/removed as a result of applying the CBTs. This implies that sample size could be a factor in effectiveness of fairness improvement by CBTs. Besides, we also observed that hardness bias tends to be associated with fairness, where an improvement on hardness bias (denoted as H-bias in Table III) of the datasets by an CBT also had improvement on fairness, e.g., all 6 CBTs which improved hardness bias also improved ABROCA in the OULA dataset. This indicated that decrease in hardness bias of a dataset could potentially help reduce the algorithmic bias.

The impact of data balancing on predictive accuracy. Compared to fairness improvement, accuracy was almost exclusively improved by over-sampling techniques, with the exception of SMOTE-TomekLink and the pipeline of top-4 class balancing techniques, in which cases under-sampling was combined with over-sampling techniques and were shown to improve accuracy. Further, the worst accuracy performance (measured by AUC) was achieved by under-sampling CBTs in all three tasks, e.g., One-Sided Selection (which combined two under-sampling techniques) had the worst AUC performance in the Moodle task, while Edited-NN had the worst AUC in KDDCUP and OULA. This indicated that over-sampling CBTs may be more preferable than under-sampling CBTs in cases where both accuracy and fairness are important in a modelling task. We also observed that accuracy improvement tended to happen simultaneously when there was fairness improvement by an over-sampling CBT, e.g., in Moodle, the accuracy-improving CBTs were also improving fairness, and in KDD-CUP, 4 out of 6 over-sampling CBTs improved both fairness and accuracy. This indicated that there was no direct trade-off relationship between fairness and accuracy, i.e., when applying CBTs to datasets, the improvement of prediction fairness for one group of students is not dependent on the sacrifice of prediction accuracy for another group of students. To further investigate the reasoning behind this, we looked into the AUC of each gender group, i.e., male and female AUC (denoted as M-AUC and F-AUC) in Figure 2. The under-sampling CBTs were omitted as they did not improve accuracy. We noted that most of the over-sampling CBTs were improving the gender group that had worse AUC (denoted as minority group) compared to the other gender group (denoted as majority group) in the baseline (i.e., no CBT applied), e.g., female AUC (F-AUC) after the application of CBTs were higher than the baseline F-AUC in KDDCUP (Figure 2(b)). Similar patterns were also observed in Moodle and OULA where the CBTs were shown to improve on the minority group AUC. However, we did not observe a consistent improvement/reduction of AUC after the application of CBTs to the majority group compared to the baseline, i.e., the majority group AUC after the application of CBTs could either slightly under- or out-perform the baseline. Therefore, under the scenario where the minority group had a better AUC (which is highly likely a result after applying over-sampling CBTs), both fairness and accuracy can be improved if the majority group AUC did not decline, or only declined marginally (i.e., the decline is not large enough to offset the gains in the minority group).

VI. Discussion

Applying modelling approaches to online education is increasingly popular in the age of big data. However, the bias of the prediction results could prevent minority group students from receiving the same high-quality teacher assistance as a result of applying the ML model. Given the wide adoption of CBTs for boosting the prediction accuracy of those modelling approaches, this study evaluated the effects of CBTs on model prediction bias between male and female students across three longstanding educational tasks.

We highlighted the following contributions of this study to the existing literature. First, we included in this study an assessment of dataset characteristics in terms of distribution and hardness bias prior to model training, among which the hardness bias were shown to be potentially associated with algorithmic unfairness. Second, we investigated the effects of a total of 11 CBTs of three types, namely, under-sampling, over-sampling, and a hybrid of these two, therefore allowing us to show the effect of CBTs on algorithmic fairness/unfairness. Third, different from previous work [19], [56], we evaluated predictive accuracy alongside fairness, and demonstrated how different CBTs affect the interplay between accuracy and fairness.

A. Implications

Firstly, we showed that all datasets had bias to a certain degree originally, two out of the three datasets had over 60% of samples in either male or female group. It also appeared that members from one gender group were more difficult to predict than members from the other. This implies that bias existed to a certain degree in real-world educational datasets, and applying such biased datasets where a demographic group was under-presented could easily lead to the model neglecting the minority group. Ideally, such bias should be addressed during the data collection stage. However, in real-world applications, this is often not possible and researchers may consider data pre-processing strategies (such as applying CBTs) to ensure parity of group representations in the dataset, as demonstrated in Table III, such CBT-balanced datasets can have lower hardness bias and may consequently lead to lower algorithmic bias.

Secondly, we found that applying CBTs to re-balance datasets was a promising approach in reducing predictive parity among different gender groups. In fact, most of the CBTs improved fairness in at least two out of the three educational tasks, and three CBTs improved fairness in all tasks. Further, the best performing CBT (i.e., Edited-NN and One-Sided Selection) reduced the predictive parity (in terms of ABROCA) by more than 50% in KDDCUP and OULA. To put it in context, KDDCUP had around 0.025 difference in AUC between male and female students based on Figure 2 and, given that large MOOC courses typically have hundreds of thousands of enrolments each year, such predictive bias could potentially affect tens of hundreds of students who were demographically under-represented (e.g., gender, race) in the datasets. These students' may be neglected by instructors while posting course related posts on the forum, or being at risk of dropout or failing the course. By applying CBTs to re-balance the dataset, the improved predictive parity shown in this study implied that students of the minority group could receive fairer attention from instructors and help reduce bias in education.

Lastly, in contrast to previous studies, which typically neglected the accuracy evaluation when evaluating fairness [56], [19]. Our study evaluated the effect of CBTs on accuracy alongside fairness and showed that fairness improvements tended to be associated with accuracy improvements when applying over-sampling CBTs. Upon digging deeper into male and female AUC scores, we found that an improved fairness was typically associated with an improved AUC score on the minority group that performed worse (in terms of AUC) than the other group when models were trained using the original datases without the application of CBTs. This implies that not only there is no trade-off between the two, but that improving fairness may potentially be complementary to accuracy. In other words, by adopting over-sampling techniques which generated more minority samples to empower the model to learn better representations of the minority group, not only the gap between the minority and majority groups became smaller, the overall accuracy was also better. Future studies could follow this line of research and try to design an CBT method that can enhance both at the same time.

B. Limitations and future research

Firstly, this study focused on student gender groups to study effectiveness of CBTs on improving bias in predictive modeling. More studies are needed on other demographic attributes, such as first-language and education background, to further validate the results reported in the current study. Secondly, unlike accuracy metrics, fairness metrics are still in an early stage of research. The state-of-art fairness metric ABROCA advanced fairness evaluation in several aspects noted in Section III-B. We highlight several gaps in the fairness research which may undermine the adoption of ABROCA: (i) interpretation of ABROCA is not as straight forward compared to interpreting accuracy metrics and future studies should aim to visualise ABROCA results to better represent the level of fairness/unfairness; (ii) more research should be conducted to enable a better understanding of the relationship between an ABROCA value and its practical impact, e.g., how many students would likely be treated unfairly by a predictive model with an ABROCA value of 0.1. Thirdly, we found that certain datasets (e.g., Moodle) received less fairness improvement from CBTs compared to other datasets. Though we speculated that this may be due to the dataset had a small number of records which could cause CBTs to operate on a limited sample size (and end up generating/removing a limited number of samples), more research is needed to verify the significance of sample size when applying CBTs. We plan on conducting experiments to evaluate CBTs' effectiveness by experimenting with training data of different sample sizes in future studies. Lastly, we acknowledged that bias may occur at any stage of the ML pipeline, and the present study addressed only the data aspect of algorithmic biases. Other biases, which may occur during such phases as the annotation of a training sample [71], feature engineering [72], and model training [69], were not tackled in this study. Future works may investigate approaches to de-bias other aspects of algorithmic bias and how such de-biasing approaches may be associated with accuracy and fairness of predictive models.

REFERENCES

[1] O. Almatrafi, A. Johri, and H. Rangwala, "Needle in a haystack: Identifying learner posts that require urgent response in mooc discussion forums," *Computers & Education*, vol. 118, pp. 1–9, 2018.

[2] A. F. Wise, Y. Cui, and J. Vytasek, "Bringing order to chaos in mooc discussion forums with content-related thread identification," in *LAK*, 2016, pp. 188–197.

[3] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, I. Estévez-Ayres, and C. D. Kloos, "Sentiment analysis in moocs: A case study," in *2018 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2018, pp. 1489–1496.

[4] W. Feng, J. Tang, and T. X. Liu, "Understanding dropouts in moocs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 517–524.

[5] E. S. Acosta and J. J. E. Otero, "Automated assessment of free text questions for mooc using regular expressions," *Information Resources Management Journal (IRMJ)*, vol. 27, no. 2, pp. 1–13, 2014.

[6] J. M. Garcia-Gorrostieta and A. López-López, "A corpus for argument analysis of academic writing: argumentative paragraph detection," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4565–4577, 2019.

[7] K. Mongkhonvanit, K. Kanopka, and D. Lang, "Deep knowledge tracing and engagement with moocs," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 2019, pp. 340–342.

[8] V. Kovanović, S. Joksimović, Z. Waters, D. Gašević, K. Kitto, M. Hatala, and G. Siemens, "Towards automated content analysis of discussion transcripts: A cognitive presence case," in *LAK*, 2016, pp. 15–24.

[9] V. Neto, V. Rolim, A. P. Cavalcanti, R. D. Lins, D. Gasevic, and R. Ferreiramello, "Automatic content analysis of online discussions for cognitive presence: A study of the generalizability across educational contexts," *IEEE Transactions on Learning Technologies*, 2021.

[10] 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.

[11] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[13] G. E. Batista, R. C. Prati, and M. C. Monard, "Balancing strategies and class overlapping," in *International Symposium on Intelligent Data Analysis*. Springer, 2005, pp. 24–35.

[14] A. P. Cavalcanti, A. Diego, R. F. Mello, K. Mangaroska, A. Nascimento, F. Freitas, and D. Gašević, "How good is my feedback? a content analysis of written feedback," in *Proceedings of the tenth international conference on learning analytics & knowledge*, 2020, pp. 428–437.

[15] G. Barbosa, R. Camelo, A. P. Cavalcanti, P. Miranda, R. F. Mello, V. Kovanović, and D. Gašević, "Towards automatic cross-language classification of cognitive presence in online discussions," in *Proceedings of the tenth international conference on learning analytics & knowledge*, 2020, pp. 605–614.

[16] B. Clavié and K. Gal, "Edubert: Pretrained deep language models for learning analytics," *arXiv preprint arXiv:1912.00690*, 2019.

[17] J. Gardner, C. Brooks, and R. Baker, "Evaluating the fairness of predictive student models through slicing analysis," in *Proceedings of the 9th international conference on learning analytics & knowledge*, 2019, pp. 225–234.

[18] S. Doroudi and E. Brunskill, "Fairer but not fair enough on the equitability of knowledge tracing," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, ser. LAK19. New York, NY, USA: Association for Computing Machinery, 2019, p. 335–339. [Online]. Available: https://doi.org/10.1145/3303772.3303838

[19] S. Yan, H.-t. Kao, and E. Ferrara, "Fair class balancing: enhancing model fairness without observing sensitive attributes," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1715–1724.

[20] P. Plaza, M. Castro, J. Merino, T. Restivo, A. Peixoto, C. Gonzalez, A. Menacho, F. García-Loro, E. Sancristobal, M. Blazquez *et al.*, "Educational robotics for all: Gender, diversity, and inclusion in steam," in *2020 IEEE Learning With MOOCS (LWMOOCS)*. IEEE, 2020, pp. 19–24.

[21] S. Jiang, K. Schenke, J. S. Eccles, D. Xu, and M. Warschauer, "Cross-national comparison of gender differences in the enrollment in and completion of science, technology, engineering, and mathematics massive open online courses," *PloS one*, vol. 13, no. 9, p. e0202463, 2018.

[22] E. Farrow, J. Moore, and D. Gašević, "Analysing discussion forum data: a replication study avoiding data contamination," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 2019, pp. 170–179.

[23] G. Barbosa, R. Camelo, A. P. Cavalcanti, P. B. C. de Miranda, R. F. Mello, V. Kovanovic, and D. Gasevic, "Towards automatic cross-language classification of cognitive presence in online discussions," in *Proceedings of the 10th International Conference on Learning Analytics and Knowledge, Frankfurt, Germany, March 23-27, 2020*, C. Rensing and H. Drachsler, Eds. ACM, 2020, pp. 605–614. [Online]. Available: https://doi.org/10.1145/3375462.3375496

[24] Y. Xu and C. F. Lynch, "What do you want? applying deep learning models to detect question topics in mooc forum posts?" in *Woodstock'18: ACM Symposium on Neural Gaze Detection*, 2018, pp. 1–6.

[25] S. Joksimović, V. Kovanović, and S. Dawson, "The journey of learning analytics," *HERDSA Review of Higher Education*, vol. 6, pp. 27–63, 2019.

[26] S. Lee and J. Y. Chung, "The machine learning-based dropout early warning system for improving the performance of dropout prediction," *Applied Sciences*, vol. 9, no. 15, p. 3093, 2019.

[27] D. R. Garrison, *Thinking collaboratively: Learning in a community of inquiry*. Routledge, 2015.

[28] Y. Cui and A. F. Wise, "Identifying content-related threads in mooc discussion forums," in *Learning@Scale*, 2015, pp. 299–303.

[29] A. Agrawal, J. Venkatraman, S. Leonard, and A. Paepcke, "Youedu: addressing confusion in mooc discussion forums by recommending instructional video clips," 2015.

[30] R. Pekrun, "The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice," *Educational psychology review*, vol. 18, no. 4, pp. 315–341, 2006.

[31] A. Barbosa, M. Ferreira, R. Ferreira Mello, R. Dueire Lins, and D. Gasevic, "The impact of automatic text translation on classification of online discussions for social and cognitive presences," in *Proceedings of the 11th International Learning Analytics and Knowledge Conference*, 2021, pp. 77–87.

[32] S. A. Geller, N. Hoernle, K. Gal, A. Segal, A. X. Zhang, D. Karger, M. T. Facciotti, and M. Igo, "# confused and beyond: detecting confusion in course forums using students' hashtags," in *LAK*, 2020, pp. 589–594.

[33] A. Bakharia, "Towards cross-domain mooc forum post classification," in *Learning@Scale*, 2016, pp. 253–256.

[34] J. Dong, F. He, Y. Guo, and H. Zhang, "A commodity review sentiment analysis based on bert-cnn model," in *2020 5th International Conference on Computer and Communication Systems (ICCCS)*. IEEE, 2020, pp. 143–147.

[35] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and smote," *Information Sciences*, vol. 465, pp. 1–20, 2018.

[36] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, "Smote-rs b*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory," *Knowledge and information systems*, vol. 33, no. 2, pp. 245–265, 2012.

[37] V. Neto, V. Rolim, R. Ferreira, V. Kovanović, D. Gašević, R. D. Lins, and R. Lins, "Automated analysis of cognitive presence in online discussions written in portuguese," in *Proceedings of the 13th European Conference on Technology Enhanced Learning*. Springer, 2018, pp. 245–261.

[38] S.-J. Yen and Y.-S. Lee, "Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset," in *Intelligent Control and Automation*. Springer, 2006, pp. 731–740.

[39] P. Hart, "The condensed nearest neighbor rule (corresp.)," *IEEE transactions on information theory*, vol. 14, no. 3, pp. 515–516, 1968.

[40] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Conference on Artificial Intelligence in Medicine in Europe*. Springer, 2001, pp. 63–66.

[41] I. Mani and I. Zhang, "knn approach to unbalanced data distributions: a case study involving information extraction," in *Proceedings of workshop on learning from imbalanced datasets*, vol. 126. ICML United States, 2003.

[42] I. Tomek *et al.*, "Two modifications of cnn." 1976.

[43] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*. Springer, 2005, pp. 878–887.

[44] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008, pp. 1322–1328.

[45] G. E. Batista, A. L. Bazzan, M. C. Monard *et al.*, "Balancing training data for automated annotation of keywords: a case study." in *WOB*, 2003, pp. 10–18.

[46] A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," *arXiv preprint arXiv:1608.06048*, 2016.

[47] M. Kubat, S. Matwin *et al.*, "Addressing the curse of imbalanced training sets: one-sided selection," in *Icml*, vol. 97. Citeseer, 1997, pp. 179–186.

[48] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409, pp. 17–26, 2017.

[49] M. A. Tahir, J. Kittler, and F. Yan, "Inverse random under sampling for class imbalance problem and its application to multi-label classification," *Pattern Recognition*, vol. 45, no. 10, pp. 3738–3750, 2012.

[50] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 3, pp. 408–421, 1972.

[51] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation," *Knowledge-Based Systems*, vol. 89, pp. 385–397, 2015.

[52] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[53] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[54] R. Yu, H. Lee, and R. F. Kizilcec, "Should college dropout prediction models include protected attributes?" in *Proceedings of the Eighth ACM Conference on Learning@ Scale*, 2021, pp. 91–100.
[55] M. Karimi-Haghighi, C. Castillo, D. Hernandez-Leo, and V. M. Oliver, "Predicting early dropout: Calibration and algorithmic fairness considerations," *arXiv preprint arXiv:2103.09068*, 2021.
[56] S. Riazy, K. Simbeck, and V. Schreck, "Fairness in learning analytics: Student at-risk prediction in virtual learning environments." in *CSEDU (1)*, 2020, pp. 15–25.
[57] O. B. Deho, C. Zhan, J. Li, J. Liu, L. Liu, and T. Duy Le, "How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics?" *British Journal of Educational Technology*, 2022.
[58] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan, "Algorithmic fairness," in *Aea papers and proceedings*, vol. 108, 2018, pp. 22–27.
[59] D. Pessach and E. Shmueli, "Algorithmic fairness," *arXiv preprint arXiv:2001.09784*, 2020.
[60] X. Wei, H. Lin, L. Yang, and Y. Yu, "A Convolution-LSTM-Based Deep Neural Network for Cross-Domain MOOC Forum Post Classification," *Information*, vol. 8, no. 3, p. 92, Sep. 2017, number: 3 Publisher: Multidisciplinary Digital Publishing Institute.
[61] S. X. Guo, X. Sun, S. X. Wang, Y. Gao, and J. Feng, "Attention-based character-word hybrid neural networks with semantic and structural information for identifying of urgent posts in mooc discussion forums," *IEEE Access*, vol. 7, pp. 120 522–120 532, 2019.
[62] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[63] X. Li, H. Zhang, Y. Ouyang, X. Zhang, and W. Rong, "A shallow bert-cnn model for sentiment analysis on moocs comments," in *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*. IEEE, 2019, pp. 1–6.
[64] X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting bert for end-to-end aspect-based sentiment analysis," *arXiv preprint arXiv:1910.00883*, 2019.
[65] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific data*, vol. 4, no. 1, pp. 1–8, 2017.
[66] M. R. Smith, T. Martinez, and C. Giraud-Carrier, "An instance level analysis of data complexity," *Machine learning*, vol. 95, no. 2, pp. 225–256, 2014.
[67] L. Feng, G. Liu, S. Luo, and S. Liu, "A transferable framework: Classification and visualization of mooc discussion threads," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 377–384.
[68] S. Hutt, M. Gardner, A. L. Duckworth, and S. K. D'Mello, "Evaluating fairness and generalizability in models predicting on-time graduation from college applications." *International Educational Data Mining Society*, 2019.
[69] R. S. Baker and A. Hawn, "Algorithmic bias in education," *International Journal of Artificial Intelligence in Education*, pp. 1–41, 2021.
[70] H. Wei, H. Li, M. Xia, Y. Wang, and H. Qu, "Predicting student performance in interactive online question pools using mouse interaction features," in *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 2020, pp. 645–654.
[71] E. Okur, S. Aslan, N. Alyuz, A. Arslan Esme, and R. S. Baker, "Role of socio-cultural differences in labeling students' affective states," in *International Conference on Artificial Intelligence in Education*. Springer, 2018, pp. 367–380.
[72] R. S. Diaz, F. Neutatz, and Z. Abedjan, "Automated feature engineering for algorithmic fairness."

Lele Sha is a third-year Ph.D. student in the Centre for Learning Analytics at Monash University. His main research interest centres on applying Machine Learning and Natural Language Processing techniques to automatically processing educational forum posts. Before starting his Ph.D., Lele worked on learning management systems projects, which were successfully deployed to production and currently offering hundreds of online courses on its interactive training platform for Australian students.

Mladen Raković is a Research Fellow at the Faculty of Information Technology, Monash University, Melbourne, VIC, Australia. His research interests span the fields of natural language processing, predictive modeling, computational linguistics, and educational psychology. He focuses on the development and evaluation of computer-based writing systems that monitor undergraduates' writing activity and generate learning analytics to promote self-regulated learning and deep engagement with disciplinary content.

Angel Das is a postgraduate student at Monash University, Faculty of IT. His main research interests centre on applying Deep Learning and Natural Language Processing techniques on unstructured data. Before starting his post graduation, Angel has worked with Fortune 500 healthcare organisations, helping them build Data Science capabilities and drive better patient outcomes through Social Media Analytics.

Dragan Gašević is Distinguished Professor of Learning Analytics in the Faculty of Information Technology and Director of the Centre for Learning Analytics at Monash University. As the past president (2015–2017) and a co-founder of the Society for Learning Analytics Research, he had the pleasure to serve as a founding program chair of the International Conference on Learning Analytics and Knowledge (LAK) and a founding editor of the Journal of Learning Analytics. His research centres on self-regulated and social learning, higher education policy, and data mining. He is a frequent keynote speaker and a (co-)author of numerous research papers and books.

Guanliang Chen is serving as a Lecturer in the Faculty of Information Technology, Monash University in Melbourne, Australia. Before joining Monash University, Guanliang obtained his Ph.D. degree at the Delft University of Technology in the Netherlands, where he focused on the research on large-scale learning analytics with a particular focus on the setting of Massive Open Online Courses. Currently, Guanliang is mainly working on applying novel language technologies to build intelligent educational applications. His research works have been published in international journals and conferences including AIED, EDM, LAK, L@S, EC-TEL, ICWSM, UMAP, Web Science, Computers & Education, and IEEE Transactions on Learning Technologies. Besides, he co-organized two international workshops and has been invited to serve as the program committee member for international conferences such as LAK, FAT, ICWL, etc.