IADIS International Journal on WWW/Internet
Vol. 24, No. 2, pp. 1-16
ISSN: 1645-7641

FAIRNESS-AWARE LEARNING ANALYTICS FOR EQUITABLE ADAPTIVE GAMIFICATION IN INCLUSIVE ONLINE EDUCATION

Olayinka Anthony Ojo[1], Owolabi Paul Adelana[2], Abosede Adebamise Abosede[1], Omolade Joy Omololu[1], Mariam Osayi Odu[1] and Elizabeth Akinola[1]
[1]*University of Greater Manchester, Bolton, United Kingdom*
[2]*The Open University, England, United Kingdom*

ABSTRACT

The need to ensure that all learners, particularly those with disabilities, learn equitably has led to increasing reliance on data-driven adaptive learning technologies (ALTs). While ALTs create opportunities for personalised learner support, they also raise concerns about whether predictive decisions operate inclusively across different groups of learners. To address this gap, this study extends fairness-aware learning analytics for adaptive gamification by exploring fairness trade-offs, predictive performance, robustness and early-warning capability among students with and without declared disabilities. The Open University Learning Analytics Dataset was used to construct an analytical dataset comprising 32,593 learner-course records, of which 3,164 represented students with a declared disability. Logistic regression and gradient boosting models were evaluated using predictive and group-level fairness measures. Gradient boosting showed stronger baseline performance, with an accuracy of 0.941, ROC-AUC of 0.984 and F1-score of 0.940. Although differences in overall positive prediction rates remained, multi-metric fairness-aware threshold selection reduced true-positive and false-positive rate disparities while maintaining predictive performance. Bootstrap analysis indicated uncertainty in some fairness differences, while repeated cross-validation affirmed the stability of the main findings. Feature-importance analysis showed that full-course prediction was strongly influenced by assessment information, whereas prediction based only on early behavioural indicators was substantially more difficult. The findings support a fairness-aware approach in which predictive accuracy, multiple fairness dimensions and intervention timing are jointly considered when informing adaptive gamification for inclusive online learning.

KEYWORDS

Adaptive Gamification, Algorithmic Fairness, Learning Analytics, Inclusive Learning, Educational Data Mining, Students with Disabilities

1. INTRODUCTION

Higher education has increasingly shifted towards online and distance learning, which provide flexible and scalable opportunities to a diverse group of students. Although such environments offer opportunities to increase participation, it is becoming more apparent that they can also reproduce or even aggravate existing inequities when learner diversity is not effectively managed. In particular, students with disabilities continue to face challenges related to accessibility, engagement, and timely academic support in the digital learning environment (Seale, 2014; Iniesto et al., 2023). The ability of online learning systems to promote inclusive participation is thus an important issue in regard to sustainable education. In this context, fairness in data-driven systems becomes a key mechanism through which sustainability in education can be achieved.

Gamification is widely accepted as one of the strategies for enhancing engagement in online learning, and it involves the use of game-inspired features including feedback, progress indicators, and rewards. Recent studies suggest that gamification has the potential to promote learner engagement and retention. However, the literature also indicates that its impact is quite diverse and highly contingent on design and context-specific characteristics of learners (Toda et al., 2021; Hamari et al., 2014). Recent literature also indicates that gamified systems have the potential to put learners with other cognitive and accessibility requirements at a disadvantage unintentionally unless adaptive mechanisms are finely tuned (Toda et al., 2023). These results lead to critical concerns regarding how gamification should be designed to support learners rather than reinforce existing inequalities.

At the same time, learning analytics has become increasingly integrated into online education, allowing institutions to study learner behaviour, anticipate academic performance, and deploy interventions. Engagement-based predictors, including early involvement and the frequency of interactions, are commonly used in predictive models and early-warning systems. Nevertheless, recent scholarship warns not to use such indicators as neutral proxies for learning as engagement behaviours are shaped by social, technical, and accessibility-related factors (Wise et al., 2021; Joksimovic et al., 2022). In the case of disabled students, variation in the patterns of engagement can be due to structural limitations and not learning capacity or interest.

The issue of bias and inequity in the educational prediction systems has become prominent in recent years. Empirical research shows that predictive models can be highly accurate in general and inconsistent in demographic groups in the absence of fairness as a key factor (Kizilcec and Lee, 2022; Baker and Hawn, 2022). Such disparities are particularly consequential in educational contexts, and predictive outputs could affect access to academic support, feedback, or intervention.

As a result of these issues, there has been an increasing interest in fairness-aware analytics that would provide balanced predictive systems amongst groups of learners. Also, recent scholarly works have suggested that fairness ought not to be considered as a post hoc solution but should form an integral part of design in education (Madaio et al., 2022; Barocas et al., 2023). It should, however, be noted that studies that incorporate fairness analytics for adaptive gamification in inclusive learning are rare.

In order to address this gap, this paper examines how fairness-aware learning analytics could foster adaptive gamification in virtual learning environments. With the help of large-scale behavioural and assessment data of the Open University Learning Analytics

Dataset (Kuzilek et al., 2017), the study examines learner engagement patterns, evaluates predictive model performance across disability groups, and applies fairness-aware threshold calibration. The study integrates analytical evidence and design considerations to support the development of inclusive, ethical and sustainable online learning systems. It is worth noting that sustainability, in this study, describes social inclusion in addition to educational systems with a long-term viability and fairness. Fairness-aware analytics can contribute to sustainable education through equitable resource allocation, reduced systemic bias over time and sustained learner participation across diverse demographic groups.

Accordingly, the aim of this study is to develop and evaluate a fairness-aware learning analytics approach for adaptive gamification that balances predictive performance with equitable treatment of learners with and without declared disabilities. The research question addressed by this study is: How can fairness-aware learning analytics support equitable adaptive gamification while maintaining reliable prediction of learner outcomes? For this question to be addressed, (i) predictive performance and fairness were evaluated across disability groups, (ii) fairness-aware threshold adjustment was examined, and (iii) model robustness, interpretability and early-warning capability were assessed.

2. LITERATURE REVIEW

Recent studies have increasingly questioned whether engagement-driven educational technologies sufficiently support diverse learners. While gamification still stands as a popular technique for advancing participation in online learning, recent reviews stress that its effects are uneven and highly sensitive to learner characteristics and design context (Toda et al., 2021; Hamari et al., 2014). Recent reviews caution that gamification may unintentionally disadvantage learners with different cognitive, accessibility, or health-related needs if adaptive mechanisms are not systematically designed (Sailer and Homner, 2020; Toda et al., 2023).

The motivational foundations of gamification continue to draw on self-determination theory, particularly the roles of autonomy, competence and relatedness in sustaining engagement (Ryan and Deci, 2000). However, current research stresses that motivation in online learning is shaped by structural and contextual factors, involving availability and institutional design, rather than by learner mood alone (Wise et al., 2021). This perspective has increased attention to inclusive learning design, particularly for students with disabilities.

Studies of accessibility and inclusive online education consistently report that students with disabilities face consistent challenges related to interface design, flexibility and cognitive load (Seale, 2014; Iniesto et al., 2023). Recent studies have shown that these challenges often produce differences in engagement patterns, complicating analytics-based interpretations of participation and success (Al-Azawei et al., 2021). Regardless, disability remains underexplored as a central variable in gamification research.

Learning analytics has emerged to support large-scale analysis of learner behaviour, with engagement indicators such as early involvement and interaction often widely used to predict academic outputs (Long and Siemens, 2011; Baker and Inventado, 2014). Nevertheless, current scholarship cautions against handling engagement metrics as neutral proxies for learning, noting that behavioural data are shaped by social, technical and accessibility-related conditions (Joksimović et al., 2022; Wise et al., 2021).

Concerns about bias and inequality in educational prediction systems have become more pronounced in recent years. Empirical studies have established that predictive models may achieve high overall accuracy while producing unequal outcomes across demographic groups when fairness is not explicitly examined (Kizilcec and Lee, 2022; Baker and Hawn, 2022). These findings have strengthened calls for responsible and transparent use of learning analytics in educational decision-making.

Fairness-aware machine learning provides a set of practical tools for tackling such inequalities. Building on earlier theoretical foundations (Hardt et al., 2016), current work investigates how fairness metrics and post-processing methods can be adopted in educational contexts to reduce group-level disparities (Madaio et al., 2022; Holstein et al., 2023). Scholars increasingly argue that fairness should be treated as a design requirement rather than a corrective measure adopted after deployment (Barocas et al., 2023; Kizilcec et al., 2023).

At the same time, research on ethical AI in education highlights the importance of aligning adaptive technologies with broader goals of inclusion and sustainability (Holmes et al., 2019; Williamson and Eynon, 2023). Although these discussions recognise the risk associated with data-driven personalization, limited empirical work has evaluated how fairness-aware analytics can inform adaptive gamification. This study addresses this gap by integrating learning analytics, fairness evaluation, and gamification within an inclusive and sustainability-oriented framework. While prior studies have explored pre-processing (data balancing) and in-processing (fairness-constrained models) approaches, this study adopts a post-processing threshold calibration method because of its practical compatibility with existing learning analytics systems.

3. METHODOLOGY

This study adopts a quantitative, data-driven methodology based on learning analytics to investigate the engagement, prediction, and fairness in online education.

3.1 Dataset and Analytical Sample

A fully anonymised dataset containing distance-learning students' virtual learning environment interactions and course assessments was obtained from the Open University Learning Analytics Dataset (OULAD). Data synthesis generated 32,593 learner-course records of which 29,429 (90.29%) represented students without declared disability and 3,164 (9.71%) represented students with a declared disability. Academic attainment was represented as a binary outcome comprising successful (15,385) and non-successful records (17,208).

3.2 Feature Engineering

Using the steps inherent in the preparation of raw data via machine learning methods (Popov, 2023), behavioural and assessment features were extracted from the combined dataset (Table 1). Full-course components included maximum and mean daily clicks, early active days, total clicks, active days, and early clicks. Others included mean assessment score, number of assessments, score and number of submissions, weighted assessment, and engagement consistency. For the early-warning analysis, only early clicks, early active days

and engagement consistency were used. This enabled this study to explore whether strong predictive performance relied on evidence produced later in a course (Conijn et al., 2017; Macfadyen and Dawson, 2010). The feature selection was intended to capture both the intensity and temporal pattern of learner participation while retaining assessment indicators of academic progress. Disability status was treated as a sensitive feature for fairness assessment rather than as a predictor of academic competence (Caton and Haas, 2024; Baker and Hawn, 2022).

Table 1. Feature sets used in the analysis

Feature group	**Variables**	**Purpose**
Full-course behavioural features	Total clicks, active days, maximum daily clicks, mean daily clicks, early clicks, early active days, consistency	Capture overall and temporal engagement
Assessment features	Total assessments, mean score, weighted score, number of submissions	Represent academic progress
Early-warning features	Early clicks, early active days, consistency	Test prediction using only early behavioural information
Protected attribute	Disability status	Group-level fairness evaluation
Outcome	Academic success	Prediction target

3.3 Predictive Modelling

In this study, logistic regression (which provided an interpretable linear benchmark) and gradient boosting (which captured non-linear relationships among engagement and assessment features) constituted the two supervised learning algorithms evaluated (Friedman, 2001; Tolles and Meurer, 2016). Prior to model fitting, continuous features used in logistic regression were standardised. The data were then divided into training and held-out test sets using stratification across academic outcomes and disability status (Schielzeth, 2010; Gelman, 2008). Predictive performance was evaluated using accuracy, ROC-AUC, precision, recall, F1-score and Brier score. Finally, principal analysis utilised the full-course feature set, while the early-warning feature set was assessed independently as a sensitivity analysis.

3.4 Fairness Evaluation and Threshold Selection

To examine fairness among students with and without declared disability, true positive rates (TPR), positive prediction rates, false positive rates (FPR), average predicted scores, and subgroup AUC were utilised. Also, group disparities were calculated as the measure for students with disabilities minus the corresponding measure for students without declared disabilities. Values nearer to zero therefore revealed greater uniformity (Kamiran and Calders, 2012; Richardson, 2009). To avoid considering one fairness measure in isolation, a constrained multi-metric threshold-selection procedure was applied. Decision thresholds from 0.01 to 0.99 were assessed by jointly considering TPR, FPR and positive-rate imbalances while constraining substantial deterioration in accuracy and F1-score. This approach differs from selecting a threshold primarily to minimise the TPR gap and enables a broader

assessment of the relationship between fairness and predictive utility (Corbett-Davies et al., 2017). Table 2 shows the fairness metrics that were used for evaluating the models at the group-level.

Table 2. Fairness measures used for model evaluation

Metric	Description	Fairness Dimension
Demographic Parity	Difference in overall positive prediction rates	Equal treatment
Equal Opportunity (TPR Gap)	Difference in true positive rates	Access equity
Predictive Equality (FPR Gap)	Difference in false positive rates	Error equity
Group AUC	Discrimination power within each subgroup	Model parity

3.5 Robustness and Sensitivity Analysis

To ensure the reliability of the findings of this study, three supplementary techniques were applied. This started with a five-fold repeated stratified cross-validation with five replications which explored the stability of predictive and fairness outcomes across different data divisions. Following this, 1,000 bootstrap resamples were applied to determine 95% confidence intervals for group-level fairness gaps. Feature importance was assessed using standardised logistic regression coefficients for LR and permutation importance for GB (Krstajic et al., 2014; Efron and Tibshirani, 1986). Also, to compare the full eleven-component models with models restricted to three early behavioural features, an early-warning sensitivity analysis was implemented. Finally, robustness and sensitivity analysis was completed with exploratory intersectional analysis which explored disability combined with age band, educational background, and gender, and where subgroup sizes were sufficient for useful descriptive comparison (Bauer et al., 2021).

4. RESULTS AND DISCUSSION

4.1 Baseline Predictive Performance

The two models revealed strong predictive performance using the full-course feature set. Gradient boosting (accuracy: 0.9411, ROC-AUC: 0.9836 and F1-score: 0.9395) performed better than logistic regression (accuracy: 0.9032, ROC-AUC: 0.9634 and F1-score: 0.9000) across the principal measures as shown in Table 3. Also, a comparison of the Brier score indicated gradient boosting (GB) had stronger probabilistic performance of 0.0445 against 0.0710 recorded by logistic regression (LR). The corresponding ROC curves in Figure 1 further illustrate the stronger discrimination achieved by GB.

Comparison of both models, however, revealed differences across disability groups indicating a positive-rate gap of -0.1088 for LR and -0.1136 for GB at the standard threshold of 0.50. Smaller TPR and FPR gaps were also observed, confirming that high predictive performance alone does not ensure equitable model behaviour.

Table 3. Baseline predictive performance of the models

Model	Accuracy	ROC-AUC	Precision	Recall	F1-score	Brier Score
Logistic Regression	0.9032	0.9634	0.8779	0.9233	0.9000	0.0710
Gradient Boosting	0.9411	0.9836	0.9125	0.9680	0.9395	0.0445

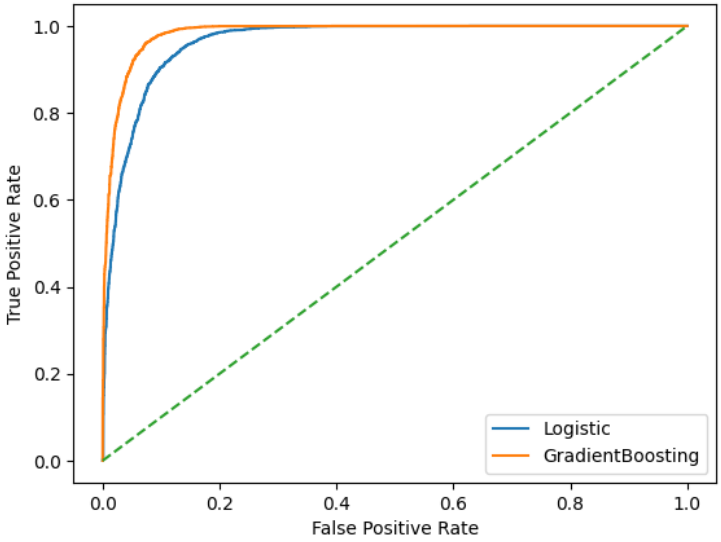

Figure 1. ROC curves for logistic regression and gradient boosting

4.2 Multi-Metric Fairness-Aware Threshold Selection

The expanded threshold analysis selected 0.66 and 0.57 for LR and GB respectively. In the case of LR, an increase in the threshold from 0.50 to 0.66 led to a reduction in the absolute TPR gap from 0.0246 to 0.0176 and the absolute FPR gap from 0.0303 to 0.0184 as shown in Table 4. Also, there was a decrease in the positive-rate gap from -0.1088 to -0.0968. This improvement involved a slight reduction in the accuracy (from 0.9032 to 0.8952) and F1-score (from 0.9000 to 0.8865).

A more favourable trade-off is provided by GB which at the selected threshold of 0.57 revealed TPR gap reduction from -0.0192 to -0.0135 and FPR gap from -0.0290 to -0.0208. However, while there was a slight decrease in positive-rate gap from -0.1136 to -0.1065, a marginal increase was observed for accuracy (0.9411 to 0.9420) and F1-score (from 0.9395 to 0.9398).

Table 4. Predictive performance and fairness before and after multi-metric threshold selection

Model	Setting	Threshold	Accuracy	F1	Positive-rate gap	TPR gap	FPR gap
Logistic Regression	Baseline	0.50	0.9032	0.9000	-0.1088	-0.0246	-0.0303
Logistic Regression	Fairness-aware	0.66	0.8952	0.8865	-0.0968	-0.0176	-0.0184
Gradient Boosting	Baseline	0.50	0.9411	0.9395	-0.1136	-0.0192	-0.0290
Gradient Boosting	Fairness-aware	0.57	0.9420	0.9398	-0.1065	-0.0135	-0.0208

The results in Table 4 therefore show that threshold selection has a varying impact on different dimensions of fairness. Improvements in error-rate parity did not eliminate disparities in overall positive prediction rates. Figures 2 – 4 further illustrate these trade-offs, showing that predictive performance remained comparatively stable across much of the threshold range, while the individual fairness measures responded differently to threshold adjustment.

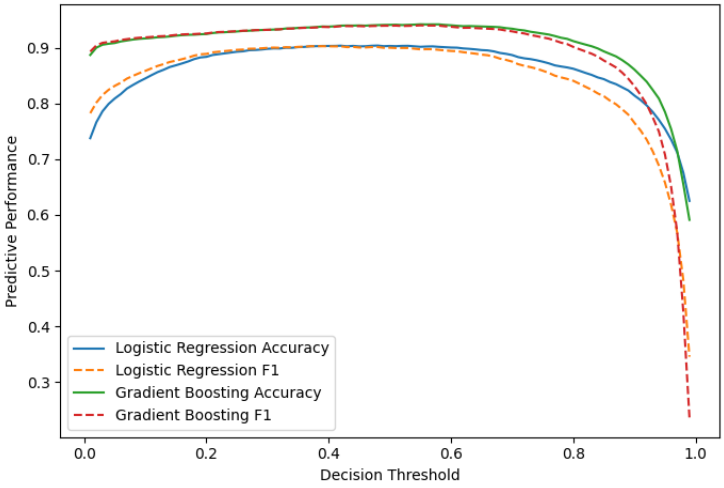

Figure 2. Predictive performance across decision thresholds

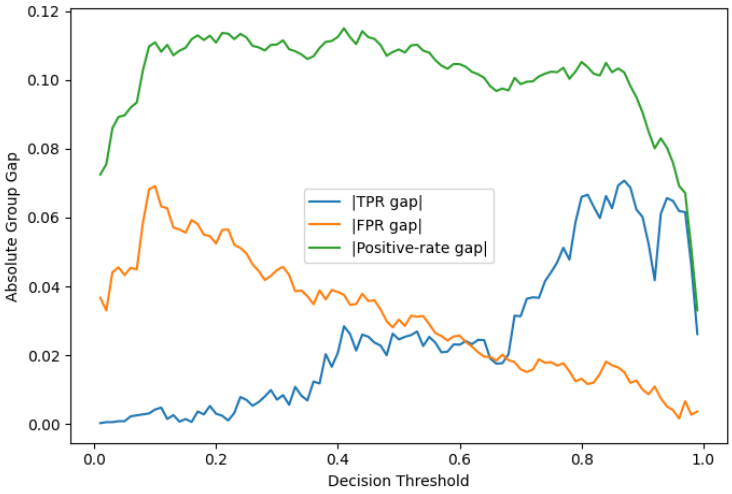

Figure 3. LR fairness trade-offs

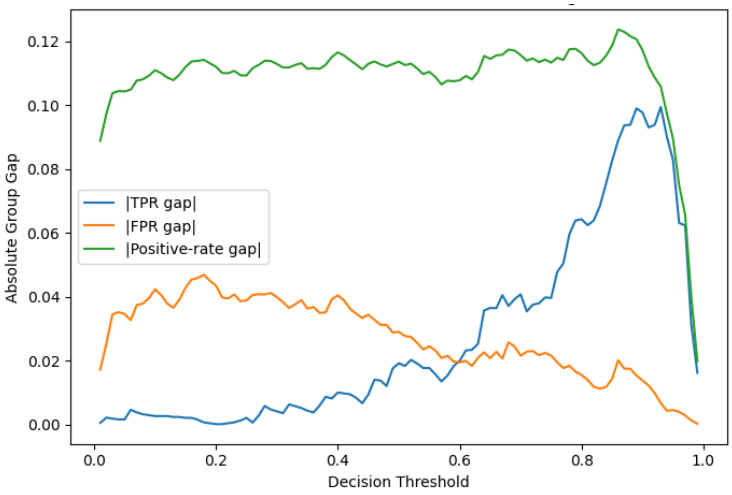

Figure 4. GB fairness trade-offs

4.3 Robustness Across Repeated Cross-Validation

Results of repeated cross-validation presented in Table 5 showed that the predictive findings were consistent. A mean accuracy of 0.9406 ± 0.0026, ROC-AUC of 0.9839 ± 0.0009 and F1-score of 0.9389 ± 0.0026 were achieved by GB. In the case of LR, mean accuracy of 0.9056 ± 0.0027, ROC-AUC of 0.9656 ± 0.0014 and F1-score of 0.9025 ± 0.0027 were obtained.

Table 5 also shows that fairness disparities were consistent across repeated folds. TPR and FPR gaps for GB were -0.0115 and -0.0208 respectively, while both gaps for LR were -0.0318 and -0.0259. The positive-rate gaps for both models were larger.

Table 5. Repeated cross-validation performance and fairness

Model	Accuracy Mean ± SD	ROC-AUC Mean ± SD	F1 Mean ± SD	TPR Gap Mean ± SD	FPR Gap Mean ± SD	Positive-Rate Gap Mean ± SD
Gradient Boosting	0.9406 ± 0.0026	0.9839 ± 0.0009	0.9389 ± 0.0026	-0.0115 ± 0.0123	-0.0208 ± 0.0145	-0.1057 ± 0.0123
Logistic Regression	0.9056 ± 0.0027	0.9656 ± 0.0014	0.9025 ± 0.0027	-0.0318 ± 0.0182	-0.0259 ± 0.0162	-0.1097 ± 0.0142

4.4 Uncertainty in Fairness Estimates

Bootstrap analysis in Table 6 provided further details about the group differences. At the fairness-aware LR threshold, the positive-rate gap was -0.0972 with a 95% confidence interval of -0.1316 to -0.0624. The corresponding gap for GB was -0.1065 with confidence interval (also below zero) of -0.1416 to -0.0719.

However, the confidence intervals for the TPR and FPR gaps moved across zero for both models. For example, the GB TPR gap was -0.0137 (95% interval of -0.0422 to 0.0107), while its FPR gap was -0.0209 with an interval of -0.0418 to 0.0023. This is an indication that the remaining error-rate differences are less certain than the persistent difference in overall positive prediction rates.

Table 6. Bootstrap 95% confidence intervals for fairness gaps

Model	Metric	Estimate	95% CI
Logistic Regression	Positive-rate gap	-0.0972	-0.1316 to -0.0624
	TPR gap	-0.0188	-0.0609 to 0.0189
	FPR gap	-0.0186	-0.0413 to 0.0051
Gradient Boosting	Positive-rate gap	-0.1065	-0.1416 to -0.0719
	TPR gap	-0.0137	-0.0422 to 0.0107
	FPR gap	-0.0209	-0.0418 to 0.0023

4.5 Feature Importance

Feature analysis showed that a combination of behavioural and assessment indicators contributed to model performance. Top features for LR were active days (2.6940), mean score (2.2433), and weighted assessment score (1.3778). For the GB model, the permutation importance shown in Figure 5 indicates that the most influential features were assessment-related, with weighted assessment score (0.1970) followed by mean assessment score (0.0761). Active days (0.0081) and total clicks (0.0001) were, however, very low. The results indicate that the very high full-course predictive performance is substantially associated with assessment information that becomes available as the course progresses.

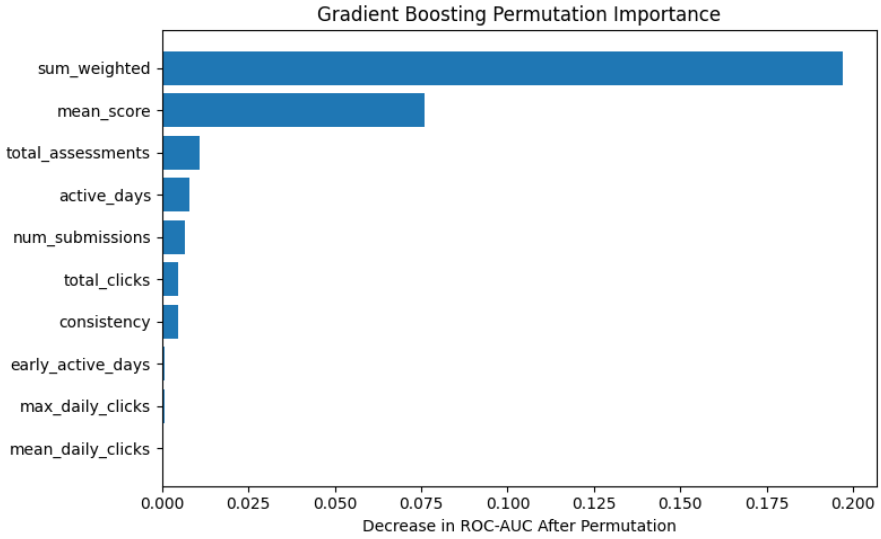

Figure 5. Gradient boosting (GB) permutation importance

4.6 Early-Warning Sensitivity Analysis

When prediction was restricted to early behavioural information, predictive performance was substantially lower. As could be seen in Table 7, early-warning LR achieved an accuracy of 0.6454, ROC-AUC of 0.7109 and F1-score of 0.5992. GB performed better, achieving an accuracy of 0.6768, ROC-AUC of 0.7573 and F1-score of 0.6715.

It is interesting to note that the early-warning models showed lower positive-rate differences than the full-course models. This improvement in some fairness measures, however, was accompanied by a considerable loss of predictive utility. The result therefore illustrates another fairness-utility trade-off: a model may display smaller group disparities while being substantially less useful for accurate early prediction.

Table 7. Full-course and early-warning model comparison

Feature regime	Model	Accuracy	ROC-AUC	F1	Positive-rate gap
Early warning	Logistic Regression	0.6454	0.7109	0.5992	-0.0174
Early warning	Gradient Boosting	0.6768	0.7573	0.6715	-0.0432
Full course	Logistic Regression	0.9032	0.9634	0.9000	-0.1088
Full course	Gradient Boosting	0.9411	0.9836	0.9395	-0.1136

4.7 Exploratory Intersectional Analysis

The exploratory analysis identified variation within the disability categories themselves. For example, among learners with declared disabilities, the females had a positive prediction rate of 0.4367 compared with 0.3582 recorded for male learners, while subgroup AUC values

stayed high at 0.9796 and 0.9935 respectively. There were also observed differences across combinations of age and educational background. Caution should be exercised when interpreting these findings because some intersectional groups were considerably smaller than others. These findings, however, indicate that a binary disability comparison can hide additional variation and support the need for more granular fairness analysis in future studies.

4.8 Fairness-Aware Adaptive Gamification Framework

A simplified framework that links predictive learning analytics with fairness assessment and adaptive gamification is presented in Figure 6. Learners' behavioural and assessment data are used to provide an estimate of academic success before model outputs are audited across disability groups. If and where disparities are identified, the decision threshold can be adjusted using multiple fairness measures while maintaining acceptable predictive performance. The resulting predictions would inform appropriate gamified support, including feedback, challenges, scaffolding and progress indicators. Continuous monitoring of learner responses and outcomes, with feedback into the system, can support ongoing improvements in both effectiveness and equity.

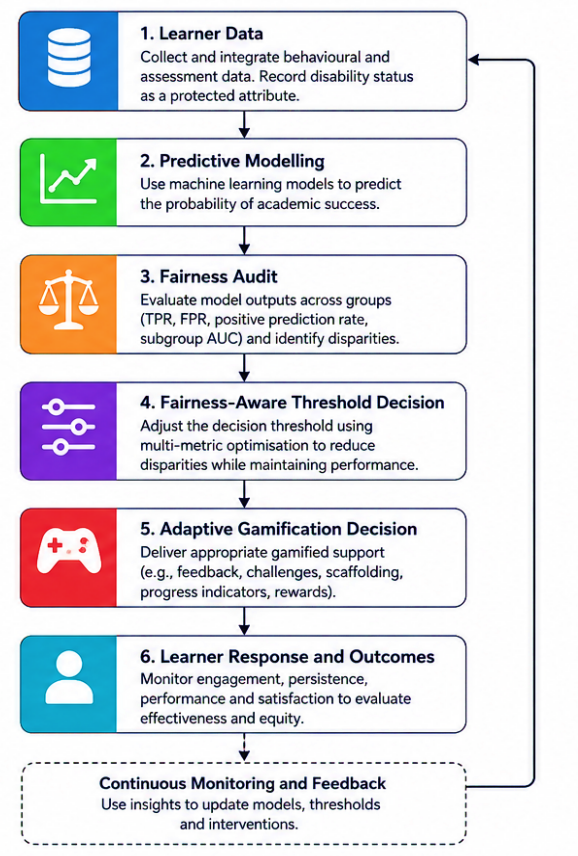

Figure 6. Fairness-aware learning analytics framework for equitable adaptive gamification

4.9 Discussion

The extended findings support the core argument that predictive accuracy and fairness should be jointly evaluated when analytics are intended to inform adaptive educational decisions. Gradient boosting provided excellent predictive performance while being stable across repeated cross-validation, although differences in positive prediction rates remained between students with and without declared disabilities. This finding is in line with concerns that educational prediction systems may perform strongly at aggregate level while retaining group-level disparities (Baker and Hawn, 2022; Kizilcec and Lee, 2022).

The multi-metric threshold analysis provides an extension to the earlier findings by showing the impossibility of reducing fairness optimisation to a single measure. When the operating threshold was adjusted, both TPR and FPR differences were reduced (especially for gradient boosting) without negatively affecting the predictive performance. However, it should be noted that the positive-rate gap was comparatively large and its bootstrap confidence interval did not include zero. It is therefore necessary that fairness-aware optimisation be interpreted as reducing particular forms of disparity rather than establishing that a model has become universally fair. This interpretation aligns with the broader view that fairness is a representation of a set of contextual (and sometimes competing) requirements rather than a single technical target (Barocas et al., 2023; Madaio et al., 2022).

An additional practical insight was provided by the feature-importance and early-warning analyses in the study. Assessment scores had a strong influence on full-course prediction, particularly in gradient boosting. However, when models were restricted to early engagement indicators, they produced only moderate discrimination. This emphasises the importance of prediction timing when fairness-aware analytics is linked with adaptive gamification. It is not possible for a system that is intended to provide early motivational support to assume the same level of predictive reliability as one operating later in the learning process. Stronger behavioural indicators or a more careful decision policy may therefore be required for early intervention.

These findings also strengthen the case for clearly distinguishing between prediction and intervention. While fairness-aware analytics can indicate whether a learner may benefit from additional support, it does not suggest that disability status should directly determine the type of gamified intervention received. An alternative is that protected characteristics should mainly support fairness auditing, while personalised feedback, scaffolding, challenges or progress mechanisms respond to individual learning states. This approach reduces the risk of replacing analytical bias with disability-based stereotyping.

Further, the exploratory intersectional results show that students recorded under the same broad disability category do not form a homogeneous group. Observed differences associated with gender, age and previous education are indicative of possible hidden complex patterns in two-group comparisons. For inclusive adaptive systems, therefore, it is recommended that fairness monitoring remain iterative and context-sensitive instead of depending on a one-off comparison between two learner categories.

Finally, when viewed from a sustainability perspective, the findings suggest that fairness-aware adaptive gamification can contribute to improved long-term educational participation by reducing the possibility of data-driven support systematically favouring particular groups. It should, however, be noted that sustainable implementation depends on the combination of predictive performance with continuous fairness auditing, appropriate intervention timing and transparent decisions about acceptable trade-offs between utility and equity.

5. CONCLUSION

This study has demonstrated that fairness-aware learning analytics can support more equitable adaptive gamification without necessarily compromising predictive performance. It was found that gradient boosting (GB) attained the best predictive performance, while multi-metric threshold adjustment lowered disparities in both TPR and FPR. Differences in positive prediction rates persisted, however. Also, robustness and early-warning analyses showed that fairness should be considered together with predictive reliability and the timing of available learner data. The proposed approach therefore provides a practical basis for integrating fairness into adaptive learning decisions. The generalisability of these findings is limited by the binary modelling of disability and the absence of an empirically deployed gamification intervention. Consequently, future research should be conducted to empirically validate the framework within real learning environments using detailed accessibility information to measure its effects on engagement, persistence, and academic success.

REFERENCES

Al-Azawei, A. et al. (2021). Towards inclusive MOOCs: bridging accessibility, usability and learner characteristics. *Educational Technology Research and Development, 69*(1), 455–482.

Baker, R. S. and Inventado, P. S. (2014). *Educational data mining and learning analytics*. Learning Analytics. Springer, New York, USA, 61–75.

Baker, R. S. and Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*, 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Barocas, S. et al. (2023). *Fairness and machine learning: limitations and opportunities*. MIT Press, Cambridge, MA, USA.

Bauer, G. R. et al. (2021). Intersectionality in quantitative research: A systematic review of its emergence and applications of theory and methods. *SSM-Population Health, 14*, 100798.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys, 56*(7), 1-38. https://doi.org/10.1145/3616865

Conijn, R. et al. (2017). Predicting Student Performance from LMS Data: A Comparison of 17 Blended Courses Using Moodle LMS. *IEEE Transactions on Learning Technologies, 10*(1), 17-29. https://doi.org/10.1109/TLT.2016.2616312

Corbett-Davies, S. et al. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797-806).

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 54-75. https://doi.org/10.1214/ss/1177013815

Friedman, J. H., (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232. https://doi.org/10.1214/aos/1013203451

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine, 27*, 2865-2873. https://doi.org/10.1002/sim.3107

Hamari, J. et al., (2014). Does gamification work? A literature review of empirical studies on gamification. *Proceedings of the 47th Hawaii International Conference on System Sciences*. Waikoloa, Hawaii, USA, pp. 3025–3034. https://doi.org/10.1109/HICSS.2014.377

Hardt, M. et al. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*.

Holmes, W. et al. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign, Boston, USA.

Holstein, K. et al. (2023). Fairness and equity in learning analytics and educational data mining. *Journal of Learning Analytics*, *10*(2), 1–8.

Iniesto, F. et al. (2023). Accessibility in online learning: challenges and opportunities for inclusive higher education. *Distance Education, 44*(3), 321–338.

Iniesto, F. and Rodrigo, C. (2024). Understanding accessibility in MOOCs: findings and recommendations for future designs. *Journal of Interactive Media in Education*, (1), 20.

Joksimović, S. et al., (2022). Beyond engagement: rethinking learner behaviour in learning analytics. *Computers & Education, 185*, 104536.

Kamiran, F. and Calders, T., (2012). Data preprocessing techniques for classification without discrimination. *Knowl Inf Syst 33*, 1–33(2012). https://doi.org/10.1007/s10115-011-0463-8

Kizilcec, R. F. et al. (2023). Algorithmic fairness and transparency in education: challenges and opportunities. *British Journal of Educational Technology, 54*(6), pp. 2239–2255.

Kizilcec, R. F. and Lee, H. (2022). Algorithmic fairness in education. *The Ethics of Artificial Intelligence in Education*. Routledge, London, UK, pp. 174–202.

Krstajic, D. et al. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal Cheminformatics 6, 10*(2014). https://doi.org/10.1186/1758-2946-6-10

Kuzilek, J. et al., (2017). Open university learning analytics dataset. *Scientific Data, 4*, 170171.

Long, P. and Siemens, G. (2011). Penetrating the fog: analytics in learning and education. *EDUCAUSE Review, 46*(5), 31–40.

Macfadyen, L. P. and Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education, 54*(2), 588-599. https://doi.org/10.1016/j.compedu.2009.09.008

Madaio, M. A. et al. (2022). Designing for fairness in educational AI systems. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul, South Korea, pp. 190–203.

Popov, A. (2023). Feature engineering methods. In *Advanced methods in biomedical signal processing and analysis* (pp. 1-29). Academic Press. https://doi.org/10.1016/B978-0-323-85955-4.00004-1

Prinsloo, P. and Slade, S. (2017). Ethics and learning analytics. *Handbook of Learning Analytics*. SoLAR, Edmonton, Canada, pp. 49–57.

Richardson, J. T. E., (2009). The academic attainment of students with disabilities in UK higher education. *Studies in Higher Education, 34*(2), 123–137. https://doi.org/10.1080/03075070802596996

Ryan, R. M. and Deci, E. L. (2000). The "what" and "why" of goal pursuits: human needs and the self-determination of behaviour. *Psychological Inquiry, 11*(4), 227–268.

Ryan, R. M. and Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective. *Contemporary Educational Psychology, 61*, 101860.

Sailer, M. and Homner, L. (2020). The gamification of learning: a meta-analysis. *Educational Psychology Review, 32*, 77–112.

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution, 1*, 103-113. https://doi.org/10.1111/j.2041-210X.2010.00012.x

Seale, J. (2014). *E-learning and disability in higher education* (2nd ed.). Routledge, London, UK.

Toda, A. M. et al. (2019). A taxonomy of game elements for gamification in educational contexts. *Proceedings of the IEEE International Conference on Advanced Learning Technologies*. Maceió, Brazil.

Toda, A. M. et al. (2021). Analysing the long-term effects of gamification on student engagement. *Computers & Education, 170*, 104222.

Toda, A. M. et al. (2023). When gamification fails: Unintended consequences of game elements in education. *Educational Technology Research and Development, 71*(4), 1987–2009.

Tolles, J. and Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. *JAMA, 316*(5), 533-534. https://doi.org/10.1001/jama.2016.7653

Williamson, B. and Eynon, R. (2023). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology, 48*(1), 1–15.

Wise, A. F. et al. (2021). Subversive learning analytics. *Proceedings of the 11th International Conference on Learning Analytics and Knowledge*. Irvine, USA, pp. 639–645.

Supplementary Table S1. Exploratory intersectional fairness results by disability, gender, age and previous education

Dimension	Intersection group	n	Positive rate	TPR	FPR	AUC	Avg. score
gender	D0_gender=F	3286	0.5307	0.9626	0.0909	0.9790	0.4927
gender	D0_gender=M	4072	0.4966	0.9756	0.0828	0.9865	0.4750
gender	D1_gender=F	403	0.4367	0.9455	0.0840	0.9796	0.4059
gender	D1_gender=M	388	0.3582	0.9562	0.0319	0.9935	0.3425
age_band	D0_age_band=0-35	5106	0.4847	0.9661	0.0817	0.9840	0.4569
age_band	D0_age_band=35-55	2195	0.5727	0.9772	0.0998	0.9811	0.5409
age_band	D0_age_band=55<=	57	0.5965	0.9429	0.0455	0.9610	0.5810
age_band	D1_age_band=0-35	590	0.3847	0.9493	0.0563	0.9875	0.3610
age_band	D1_age_band=35-55	201	0.4378	0.9529	0.0603	0.9869	0.4153
highest_education	D0_highest_education=A Level or Equivalent	3079	0.5525	0.9726	0.0956	0.9838	0.5225
highest_education	D0_highest_education=HE Qualification	1202	0.5832	0.9866	0.0734	0.9894	0.5583
highest_education	D0_highest_education=Lower Than A Level	2925	0.4431	0.9575	0.0847	0.9779	0.4127
highest_education	D0_highest_education=No Formal quals	88	0.2955	0.8929	0.0167	0.9964	0.2801
highest_education	D0_highest_education=Post Graduate Qualification	64	0.6563	0.9756	0.0870	0.9958	0.6496
highest_education	D1_highest_education=A Level or Equivalent	380	0.4421	0.9576	0.0465	0.9889	0.4148
highest_education	D1_highest_education=Lower Than A Level	368	0.3397	0.9397	0.0635	0.9860	0.3206