International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(1s) 2026 pp. 1-19
https://doi.org/10.70917/ijcisim-2026-2092
Article

Causal eXplainable AI for Student Dropout Prediction: Detecting Proxy Bias and Causal Mismatch

Dongsoo Moon[1], Gwangsu Lee*[2]

[1]Department of Computer Education, Sungkyunkwan University, Republic of Korea (m1d2s3@gmail.com)
2School of Computer Software, Daegu Catholic University, Republic of Korea (gslee@cu.ac.kr)
***Corresponding Author:** Gwangsu Lee (gslee@cu.ac.kr)

Abstract: Although machine learning-based student dropout prediction models achieve high predictive performance, the socioeconomic biases and decision-making mechanisms latent within these black-box models remain insufficiently understood. Conventional eXplainable artificial intelligence (XAI) techniques rely on correlation-based global feature importance, which limits their ability to identify the true causal factors that substantively flip predictions. Therefore, this study proposes a "Causal Auditing" framework to quantitatively identify and correct biases. Utilizing a Portuguese university student dataset and a single ensemble baseline model (LightGBM), we generated hierarchical multivariate counterfactual scenarios that combined uncontrollable sensitive attributes (parents' background) with actionable variables (academic achievement and debt status). The results revealed a distinct "Causal Mismatch" phenomenon: although the correlation-based SHapley Additive exPlanations analysis assigned the highest importance to past first-semester grades, the core causal triggers that flipped the model decision from "Dropout" to "Graduate" were found to be second-semester curricular units and financial status. Notably, instead of forcing extreme high-risk students to achieve unreasonable grade improvements, merely resolving their debt issues and neutralizing the structural penalties associated with their parental backgrounds—when combined with minimal academic support (completing +1 additional course)—successfully flipped the predictions for the entire target cohort. These findings demonstrate that AI models adopt students' environmental disadvantages as a form of proxy bias (i.e., a "structural penalty"). The proposed framework provides a systematic methodology for verifying the transparency and causal fairness of AI in the educational domain, and contributes to the formulation of data-driven, hybrid student support policies in university administrations.

Keywords: algorithmic fairness, causal auditing framework, causal understanding, counterfactual intervention, causal mismatch, explainable AI, proxy bias

1. Introduction

The recent integration of artificial intelligence (AI) and machine learning (ML) into educational environments has driven significant advancements in predicting and managing student dropout rates [1], [2]. Early identification of at-risk students enables educational institutions to proactively intervene and allocate resources efficiently, thereby improving student retention and academic success [3]. However, as predictive models become increasingly complex to achieve higher accuracy, their decision-making processes often devolve into "black boxes," rendering it difficult for educators and administrators to understand the rationale behind specific predictions [4], [5].

To overcome this lack of transparency, eXplainable AI (XAI) techniques, particularly SHapley Additive exPlanations (SHAP), have been widely adopted to interpret complex models by calculating the contribution of each feature [6], [7]. Although conventional XAI methods provide a global and local understanding of feature importance, they fundamentally rely on correlation rather than causation [8]. Consequently, these methods often assign high importance to unchangeable or sensitive attributes, such as parental background, socioeconomic status, or past academic records, simply because they are strongly correlated with the target variable [9], [10].

This correlation-based interpretation introduces a critical vulnerability in educational AI systems: structural proxy bias. When a, model heavily relies on uncontrollable environmental factors to predict student dropout, it implicitly penalizes students for their background, utilizing these sensitive attributes as proxies for academic failure [11]. Furthermore, providing explanations based on these unchangeable features offers no practical value for intervention. For example, informing a student that their high dropout risk is due to their parents' educational level provides no actionable guidance or recourse for improvement [12], [13].

This study defines this phenomenon as a causal mismatch, i.e., a discrepancy where variables identified as highly important by conventional XAI (e.g., SHAP) do not align with the actionable triggers that can actually alter the prediction outcome [14]. To resolve this mismatch and achieve genuine algorithmic fairness, the analytical paradigm should shift from passive, correlation-based observation to active, causal intervention [15].

To address these limitations, this study proposes a Causal XAI-based Fairness Auditing Framework utilizing multivariate counterfactual intervention. Unlike traditional approaches that evaluate fairness merely by masking sensitive variables, the proposed framework actively probes the model using causal counterfactual interventions to uncover hidden causal relationships and neutralize proxy biases. By fixing unchangeable sensitive attributes (e.g., parents' background) and dynamically searching for optimal combinations of actionable variables (e.g., curricular adjustments, debt resolution), the framework identifies the minimal intervention path required to flip a high-risk dropout prediction to a positive outcome (graduation).

The primary contributions of this study are as follows. First, we empirically demonstrate the structural limitations of conventional correlation-based XAI by revealing the discrepancy between SHAP-derived feature importance and actual causal triggers. Second, we propose a hierarchical causal intervention methodology that isolates actionable variables from immutable sensitive attributes, providing a mathematically grounded approach to finding optimal recourse paths. Third, we introduce a systemic auditing framework that not only evaluates algorithmic fairness through novel metrics, such as the flip success rate (FSR) and average risk reduction, but also provides concrete, actionable guidelines (e.g., completing one additional course) to neutralize structural penalties imposed on disadvantaged students.

The remainder of this paper is organized as follows. Section II reviews related work on educational data mining, XAI, and algorithmic fairness. Section III describes the proposed causal XAI-based framework and its mathematical formulation. Section IV presents the experimental setup, data preparation, and empirical results. Section IV presents the experimental setup, data preparation, and empirical results, and discusses the implications of causal mismatch and proxy bias neutralization. Finally, Section V concludes the study with directions for future research.

2. Related Work

2.1. XaiiIn Educational Prediction: Trends And Limitations

Recently, in the fields of educational data mining (EDM) and learning analytics, machine learning-based studies have been actively conducted to enable early prediction of student academic achievement and dropout risk. In particular, student dropout prediction is recognized as a core task that extends beyond mere performance assessment, as it directly supports early intervention, targeted learning assistance, and the efficient allocation of educational resources. Accordingly, the demand for the interpretability and practical applicability of prediction results has simultaneously increased [1], [5]–[9].

Amidst this trend, XAI techniques—such as SHAP, Local Interpretable Model-agnostic Explanations (LIME),, permutation importance, and partial dependence plots (PDP)—are being actively integrated into research on student performance and dropout prediction. Existing studies have demonstrated through XAI that variables related to academic achievement, enrollment and course completion, financial status, scholarship availability, and certain demographic backgrounds significantly influence prediction outcomes [1], [5]–[7], [10], [11]. However, recent studies also point out that explanation results can vary depending on the dataset, cohort, and learning context, highlighting that global feature importance does not always align with the actual predictive changes in individual cases [5], [11].

Furthermore, although XAI research in the educational domain is generally useful for identifying "which variables are important," it inherently faces limitations in causally explaining "why such results occur." For example, although SHAP and LIME provide statistical contributions to a prediction, they do not directly guarantee or confirm whether a specific variable is the actual causal trigger that flips the prediction outcome. Therefore, although conventional XAI serves as a critical diagnostic tool in educational data, it is not considered sufficient on its own to fully explain a model's actual decision-making structure and algorithmic fairness [3], [5], [12], [13].

Recently, AI research in higher education has expanded beyond the simple maximization of predictive performance. Increasing attention is now being directed toward systematic mapping and qualitative evaluation studies that ensure model transparency and fairness across the student lifecycle, from admissions to dropout prediction [16]–[20]. Nevertheless, these studies still exhibit significant limitations in quantitatively isolating and evaluating causal mechanisms within complex environments that are deeply entangled with proxy biases.

2.2. Fairness under Unawareness and Proxy Bias

One of the predominant approaches to designing fair predictive models is the "fairness through unawareness" strategy, which entails the exclusion of sensitive attributes. This approach aims to prevent the model from learning direct discrimination by omitting sensitive variables, such as gender, race, and socioeconomic background, during the training phase. However, recent research on algorithmic fairness indicates that this method may be insufficient for fully eradicating actual biases [9].

In particular, even when sensitive variables are removed, if other variables act as proxy features replacing them, the model can still indirectly learn discriminatory patterns. Cornacchia et al. demonstrated through a counterfactual reasoning-based fairness auditing framework that discriminatory judgments could be reproduced solely through correlated non-sensitive variables, even without the direct use of sensitive attributes [2]. This critical issue has been repeatedly raised in recent studies on fairness and actionable recourse, emphasizing that merely eliminating variables was inadequate for resolving structural bias [14].

Educational data inherently exacerbate this complexity. A student's academic achievement is intricately linked not only to individual effort and ability but also to multifaceted factors such as parental education, parental occupation, financial status, tuition payment status, and scholarship receipt. Consequently, even if parents' education and occupation are designated as sensitive variables and removed, features such as academic grades, the number of enrolled courses, registration-related variables, and financial indicators can act as proxies that indirectly reflect socioeconomic background. In such cases, although the predictive model may appear superficially fair, it potentially perpetuates structural disadvantages against specific demographic groups [2], [6], [9], [10].

2.3 Counterfactual Reasoning and Causal Xai

To mitigate these limitations, counterfactual reasoning and causal XAI have recently garnered significant attention. Counterfactual reasoning explores the decision boundaries of a model by posing the question "*If the value of a specific variable had been different, would the prediction outcome have changed?*", thereby enabling the analysis of potential predictive shifts at the individual instance level [2], [3], [13]. In particular, counterfactual explanations constitute a cornerstone of explainability research, as they provide human-intuitive explanations and identify the minimal perturbations that most strongly influence model predictions [3], [13].

However, counterfactual explanations do not automatically equate to causal understanding. Baron pointed out that although conventional XAI and standard counterfactual explanations could provide interpretations, they did not inherently guarantee causal understanding. In other words, the fact that a change in a variable alters the prediction value does not definitively establish that variable as a causally sufficient explanatory factor; integration with a more rigorous causal interpretation framework is required [3], [8]. Similarly, Chou et al., through a systematic review of counterfactuals and causability, concluded that although the counterfactual approach in XAI was highly promising, theoretical and methodological caution was necessary when interpreting actual causality [12].

Recently, efforts have been made to redefine counterfactual reasoning itself from the perspectives of fairness auditing and actionable recourse. For instance, Bynum et al. proposed a novel counterfactual reasoning paradigm that considered alternative initial conditions and structures, advancing beyond the traditional sensitive-attribute-interventional counterfactual fairness [13]. This discourse suggests the necessity of moving beyond simple feature importance comparisons to a methodology that verifies which specific combinations of variables actually flip the prediction outcome, subsequently contrasting these findings with XAI-derived feature importance rankings [3], [13], [14].

2.4 Research Gap and Positioning of Present Study

Synthesizing the existing literature reveals three primary research gaps. First, although XAI research in the educational domain is actively progressing, the majority of studies remain confined to correlation-based explanations—such as SHAP, LIME, and permutation importance—and fail to adequately verify the relationships with causal variables that actually induce prediction flips [1], [5], [6], [10]–[12]. Second, fairness through unawareness focuses on the mere removal of sensitive variables; it structurally fails to address the potential for biases to be perpetuated via proxy features [2], [14]. Third, despite the growing body of fairness research utilizing counterfactual reasoning, integrated studies that execute multivariate counterfactual analysis and comparative XAI validation in environments where academic and socioeconomic variables are complexly intertwined, such as educational data, remain scarce [2], [3], [13], [14].

Conventional univariate-centric counterfactual analyses and correlation-based XAI methods possess fundamental limitations in reflecting the true causal dynamics of educational environments characterized by intricately entangled variables. To overcome these limitations, this study proposes a "Causal XAI-based Fairness Auditing Framework" that integrates multivariate counterfactual reasoning and causal validation, leveraging a dataset of Portuguese university students.

Specifically, advancing beyond the traditional approach of independently manipulating individual variables, this framework simultaneously explores uncontrollable sensitive attributes (e.g., parents' education and occupation) and practically actionable variables (e.g., academic and financial status) to identify the Counterfactual Flip (Cflip), i.e., the minimal combination of perturbations required to reverse the prediction outcome.

Subsequently, by contrasting the derived actual causal triggers with SHAP-based global feature importance, the framework rigorously validates the model decision-making rationale from the perspectives of local and global causal understanding, as emphasized by Baron [8].

Consequently, this integrated approach provides an empirical framework to quantitatively diagnose the extent to which AI predictive models rely on socioeconomic proxy variables to render discriminatory judgments [2]–[4], [9], [12], [13].

3. Methodology

This section describes the causal XAI-based framework proposed to analyze the fairness of student dropout prediction models. The study comprises several methodological phases: predictive model construction, candidate feature selection, counterfactual data generation, counterfactual flip analysis, and XAI-based causal validation.

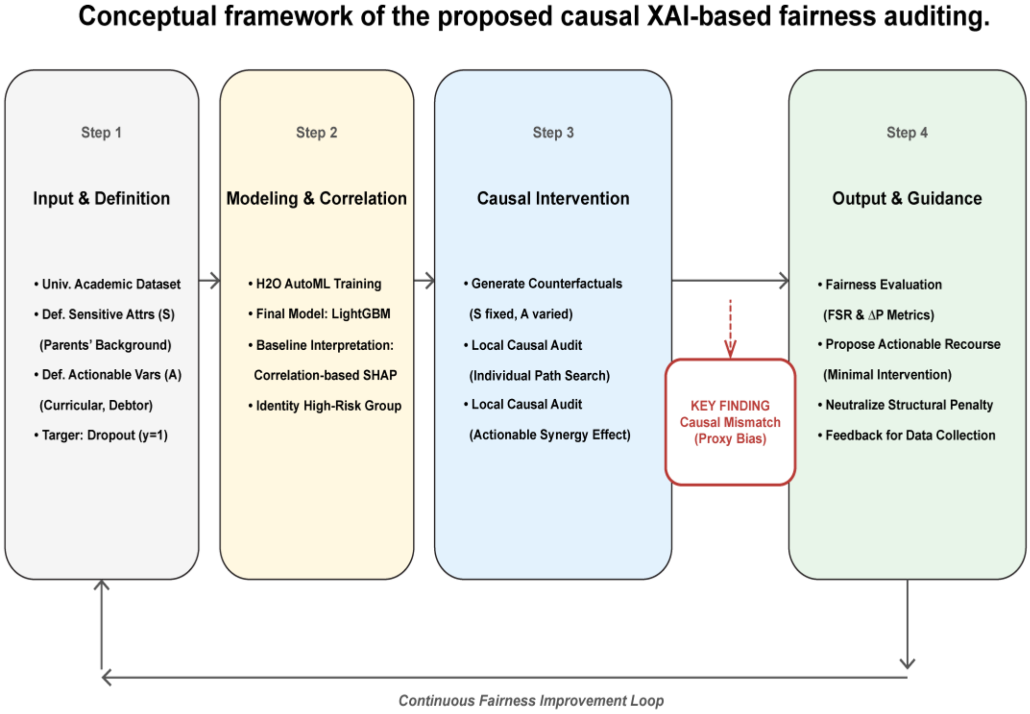

Figure 1. Conceptual framework of proposed causal XAI-based fairness auditing.

The framework is structured into a logical sequence encompassing three main phases. First, it raises the problem of "structural proxy bias," wherein conventional black-box predictive models relying on simple correlations assign disproportionately high importance to uncontrollable variables (The Problem). Second, as a resolution, it proposes a "multivariate counterfactual intervention" that combines uncontrollable sensitive attributes with actionable variables (Proposed Approach). Finally, through this causal intervention, the framework identifies the model's "causal mismatch" and neutralizes the proxy bias, thereby presenting the ultimate goal of achieving practical fairness (Ultimate Goal). (The specific data processing and modeling procedures to implement this conceptual diagram are presented in Fig. 1.)

3.1 Data Preparation

This study utilized a student dropout prediction dataset collected from undergraduate students at the Polytechnic Institute of Portalegre in Portugal, spanning the academic years 2008/2009 to 2018/2019 [7]. This dataset encompasses demographic characteristics at enrollment, socioeconomic backgrounds, and macroeconomic data. Notably, with a dropout rate reaching approximately 32%, the dataset possesses optimal characteristics to empirically demonstrate the necessity of student management and early intervention.

Table I. Summary of variables in this study and their analytical roles

Variable Group	Variables	Type	Analytical Role in This Study
Demographic data	Marital status, Nationality, Displaced, Gender, Age at enrollment, International	Categorical (nominal-coded), Binary, Discrete	General background predictors
Socioeconomic background	Mother's qualification, Father's qualification	Categorical (ordinal-coded)	Sensitive socioeconomic attributes
Socioeconomic background	Mother's occupation, Father's occupation	Categorical (nominal-coded)	Sensitive socioeconomic attributes

Socioeconomic and financial status	Educational special needs, Debtor, Tuition fees up to date, Scholarship holder	Binary	Financial and social-context predictors; potential proxy-related variables
Macroeconomic context	Unemployment rate, Inflation rate, GDP	Continuous	Contextual predictors reflecting external socioeconomic conditions
Academic background at enrollment	Application mode, Application order, Course, Daytime/evening attendance, Previous qualification, Admission grade	Categorical (nominal-coded), Ordinal/Discrete, Binary, Continuous	Baseline academic-entry predictors
Academic performance: 1st semester	Curricular units 1st sem (credited), Curricular units 1st sem (enrolled), Curricular units 1st sem (evaluations), Curricular units 1st sem (approved), Curricular units 1st sem (grade), Curricular units 1st sem (without evaluations)	Discrete count, Continuous	Core academic performance predictors
Academic performance: 2nd semester	Curricular units 2nd sem (credited), Curricular units 2nd sem (enrolled), Curricular units 2nd sem (evaluations), Curricular units 2nd sem (approved), Curricular units 2nd sem (grade), Curricular units 2nd sem (without evaluations)	Discrete count, Continuous	Core academic performance predictors
Target variable	Dropout / Graduate / Enrolled	Categorical	Prediction outcome

Table I summarizes the variables utilized in this study, categorized by feature group, data type, and analytical role. In particular, parents' educational levels and occupations were defined as sensitive socioeconomic attributes, while academic and financial variables were concurrently evaluated for their potential to act as proxy variables during the counterfactual fairness analysis. The dataset initially comprised 4,424 records, with the target variable classified into three categories representing the students' final academic status: "Graduate," "Dropout," and "Enrolled". To conduct a distinct binary classification for dropout prediction, 794 students (17.9%) labeled as "Enrolled," whose final academic status had not yet been determined, were excluded during the preprocessing stage [7]. Consequently, 3,630 records—comprising 2,209 graduates and 1,421 dropouts—were utilized for the final analysis. The detailed distribution of the target variable in the dataset is presented in Table II.

Table II. Distribution of target variables in the original dataset and final analysis

Category	Original Dataset	Preprocessed Dataset
Dropout	1,421 (32.1 %)	1,421 (39.1 %)
Graduate	2,209 (49.9 %)	2,209 (60.9 %)
Enrolled	794 (17.9 %)	- (Excluded)
Total	4,424 (100.0 %)	3,630 (100.0 %)

In this study, the following four variables were strictly defined as socioeconomic sensitive attributes:

- Mother's qualification
- Father's qualification
- Mother's occupation
- Father's occupation

Furthermore, we aimed to analyze the extent to which the prediction flip rate increased when these sensitive variables (parents' background) were strategically combined with actionable variables, specifically debt status ("Debtor") and academic achievement improvement ("Curricular units 2nd sem (approved)"). Ultimately, this analysis sought to determine whether merely removing or controlling for sensitive variables was sufficient to rectify algorithmic bias, or if active intervention was required to alter the prediction outcome and provide actionable recourse for at-risk students.

3.2 Predictive Modeling via AutoML

To predict whether a student would drop out, the H2O AutoML framework was utilized. Various machine and deep learning models were trained based on 5-fold cross-validation, and the optimal model was automatically searched. The primary algorithms employed were as follows:

Gradient boosting machine (GBM)

- Random forest (RF)
- XGBoost
- Deep neural networks
- Stacked ensemble

In this study, the H2O AutoML framework was applied to achieve optimal predictive performance, with the primary hyperparameter settings outlined in the subsequent algorithm. To ensure the reproducibility of the experiments, the random seed was fixed at 42. Considering the efficiency of the search space and computational resources, the maximum number of models to be explored was limited to 20 (max_models = 20). Furthermore, to prevent overfitting and reliably evaluate the generalization performance of the models, a 5-fold cross-validation approach was applied. In particular, addressing the class imbalance problem inherent in the dataset of this study, the area under the precision–recall curve (AUCPR) was adopted as the sorting metric for the final performance evaluation and model ranking.

Table III (Algorithm 1) presents the core modeling strategy of this study, which advances beyond a simple algorithm comparison to focus on a global exploration of the hyperparameter search space for each candidate model. The specific modeling process was designed based on the following technical validities:

- Global hyperparameter optimization (Lines 5–6): For tree-based models such as GBM and XGBoost, a random grid search space was established to yield optimal estimates with learning rates of [0.01, 0.1] and maximum tree depths of [3, 10]. For deep learning architectures, the maximum number of epochs was set to 1,000; however, an early stopping mechanism was applied to immediately halt training if the validation error did not improve for five consecutive iterations, thereby simultaneously ensuring the model convergence point and preventing overfitting.
- Cross-validation and imbalanced data handling (Lines 7–8): As AUCPR concurrently considers the precision and recall for predicting the minority class ("Dropout"), it provides a more stringent evaluation criterion than standard accuracy or ROC-AUC in an imbalanced environment.
- Final model selection considering interpretability (Lines 10–12): The final baseline model was selected as the highest-ranking single model; stacked ensembles were explicitly excluded. Although complex ensemble models exhibit superior predictive performance, their structures are excessively opaque for tracking the independent contributions of individual variables and causal changes resulting from counterfactual scenarios. Therefore, a single ensemble model (e.g., GBM or LightGBM) that learned non-linear relationships while providing highly reliable interpretations during the causal auditing process using SHAP was determined to be the optimal choice.

Table III. Algorithm 1. Pseudocode for predictive model training process

Procedure:
1: Initialize random state with seed
2: Initialize AutoML_Engine with (max_models, K, seed, metric)
3: Execute AutoML_Engine.Train(X, y, D_train)
4: For each model architecture (GBM, XGBoost, DeepLearning, GLM) do:
5: Define Hyperparameter Search Space (S):
- Learning_Rate: range (0.01, 0.1)
- Max_Depth: range (3, 10)
- Epochs/Iterations: min 10, max 1000 with Early Stopping
6: Perform Random Grid Search to find optimal parameters in S
7: Train the model on D_train using K-fold cross-validation
8: Calculate the primary metric (AUCPR)
9: End For
10: L ← Sort all evaluated models by metric in descending order
11: M_best ← Top model in L (often a Stacked Ensemble)
12: Return L, M_best

3.3 Hierarchical Causal Intervention: Multivariate Counterfactual Scenarios

To evaluate the causal robustness of the predictive model and resolve the potential biases inherent in the system, this study introduces a hierarchical causal intervention methodology based on Baron's causal understanding framework [8]. Baron emphasized that the statistical importance provided by existing XAI is insufficient to fully understand the model decision-making process; true causal mechanisms can only be identified through a counterfactual approach that explores how the prediction outcome is flipped when specific variables are artificially manipulated (intervened). Based on this theoretical philosophy, this study goes beyond merely listing which variables are important and designs a multivariate counterfactual scenario that flips the prediction through actual intervention, as follows.

If the entire feature space is denoted as X, the best model predicting a student's dropout probability is defined as (\ in [0, 1]). For a specific student instance classified into the high-risk dropout group ( ), the objective is to identify the Counterfactual Flip (CFlip) point that reverses the prediction outcome. To avoid a computationally expensive and exhaustive search, the variables are categorized based on their causal manipulability and policy effectiveness, defining the counterfactual search space hierarchically as follows:

- Socioeconomic proxy feature group (): Inherited environmental variables, such as parents' backgrounds (mother's and father's qualifications and occupations), denoted as { }These act as structural penalty factors that are impossible to alter through short-term interventions.
- Actionable feature group (): Variables that can be modified through institutional support from the university or student's individual efforts. In this study, we further subdivide this into two sub-scenarios to analyze the synergistic effect of academic intervention (number of approved curricular units in the 2nd semester) and financial intervention (resolution of debtor status).
- Based on these hierarchical feature groups, the optimization problem for finding the optimal counterfactual instance that flips the prediction outcome from y=1 (Dropout) to y=0 (Graduate) is formulated as follows:

Here, denotes a distance metric such as Gower's distance or L1/L2 norm, and τ represents the decision threshold, typically set to 0.5. This constraint-based optimization does not merely alter the prediction outcome; rather, it serves as a core mechanism to ensure causal validity by asking: "Can the student be prevented from dropping out with the minimal realistically actionable effort (academic/financial support) while leaving uncontrollable variables (e.g., parents' background) unchanged?".

3.4 Evaluation Metrics for Causal Auditing

To quantitatively verify the practical impact of the proposed multivariate causal intervention on the entire high-risk cohort (N = 128), this study introduces the following two key evaluation metrics.

- FSR: This refers to the proportion of students, initially classified in the high dropout risk group, who are successfully reclassified into the retention predicted group after applying the multivariate counterfactual intervention. It serves as a macroscopic metric that measures whether the intervention scenario can actually alter the model final decision-making outcome.

Here, N denotes the total number of high-risk students, and τ represents the decision threshold (typically 0.5). I(.) is an indicator function that returns 1 if the condition inside the parentheses (i.e., the predicted probability dropping below the threshold) is true, and 0 if it is false.

- Average risk reduction (ΔP): This represents the average decrease in the predicted dropout probability induced by the counterfactual feature perturbation (→). Although FSR is a binary metric that evaluates whether the decision boundary is crossed, ΔP measures the continuous change in probability values, thereby providing a more granular assessment of the intervention intensity and effectiveness.

Here, represents the predicted dropout probability in the original state prior to intervention, and denotes the predicted probability after the counterfactual perturbation is applied. A larger negative value for ΔP indicates a more successful reduction in the predicted dropout risk

3. 5 Evaluation Metrics for Causal Auditing

To quantitatively verify the practical impact of the proposed multivariate causal intervention on the entire high-risk cohort (N = 128), this study introduces the following two key evaluation metrics.

- FSR: This refers to the proportion of students, initially classified in the high dropout risk group, who are successfully reclassified into the retention predicted group after applying the multivariate counterfactual intervention. It serves as a macroscopic metric that measures whether the intervention scenario can actually alter the model final decision-making outcome.

Here, N denotes the total number of high-risk students, and τ represents the decision threshold (typically 0.5). I(.) is an indicator function that returns 1 if the condition inside the parentheses (i.e., the predicted probability dropping below the threshold) is true, and 0 if it is false.

- Average risk reduction (ΔP): This represents the average decrease in the predicted dropout probability induced by the counterfactual feature perturbation (→). Although FSR is a binary metric that evaluates whether the decision boundary is crossed, ΔP measures the continuous change in probability values, thereby providing a more granular assessment of the intervention intensity and effectiveness.

Here, represents the predicted dropout probability in the original state prior to intervention, and denotes the predicted probability after the counterfactual perturbation is applied. A larger negative value for ΔP indicates a more successful reduction in the predicted dropout risk.

3.6 Overall Framework

The overall workflow of this study is illustrated in Figure 2. The figure presents the detailed algorithmic flow of the causal validation process, which constitutes the core contribution of this research. Based on the trained predictive model, two parallel analytical pathways are executed. The specific procedures outlined in Figure. 2 are as follows:

(1) Baseline Predictive Model (High-Risk Identification): Based on the dropout probabilities calculated by the predictive model, students whose risk scores exceed a predefined threshold are identified as the high-risk group, thereby establishing them as the primary targets for subsequent fairness auditing.

(2) Parallel Feature Auditing: Two independent analytical pathways are conducted simultaneously on the identified high-risk cohort.

- Correlation-based XAI: The left pathway utilizes global and local SHAP techniques to extract , defining the top-k feature set that exhibits the highest statistical contribution to the prediction outcomes.
- Causal intervention: The right pathway applies multivariate counterfactual scenarios to identify, i.e., the set of decisive causal features that actually induce a decision flip in the model prediction.

(3) Causal Mismatch Calculation & Proxy Bias Detection: Finally, by comparing the variables derived from these two distinct pathways, the framework calculates the structural discrepancy (Causal Mismatch) between the derived feature importance (Global SHAP) and actual causal contribution (ΔSHAP). Within this framework, the mismatch is detected in two forms. The first is the identification of "Overestimated Features," which the model heavily relies on for decision making but possess no actual causal effect. The second is the discovery of "Hidden Causal Triggers," which are overlooked in conventional correlation-based explanations but actually flip the prediction outcome upon multivariate intervention. Consequently, this process serves to transparently audit the socioeconomic proxy bias latent within the black-box model. The ultimate fairness and intervention effectiveness of the system are rigorously evaluated using the FSR and ΔP defined in Section III-D.

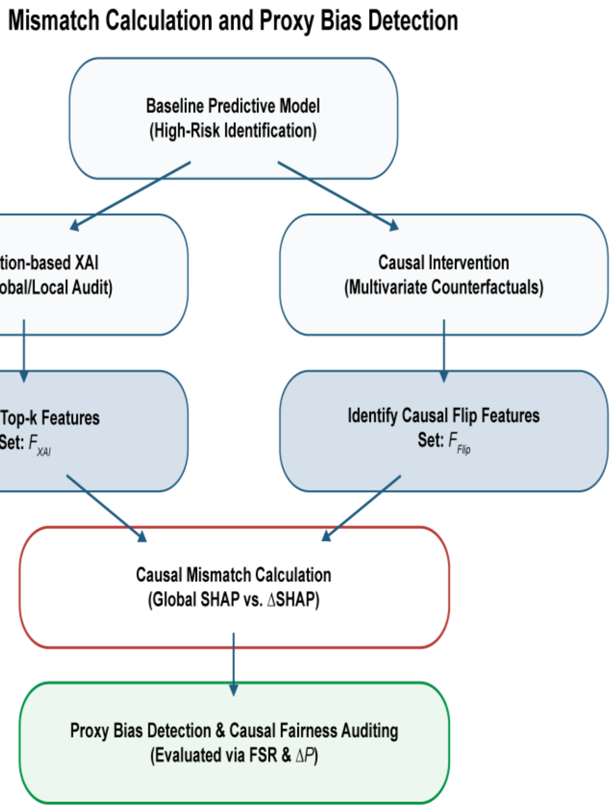

Figure 2. Flowchart of mismatch calculation and proxy bias detection

4. Results and Discussion

4.1 Predictive Model Evaluation and Interpretable Best Model Selection

Prior to applying the causal intervention framework, which was the core of this study, a baseline predictive model selection process was conducted to accurately identify the cohort of students at a high risk of dropping out.

The H2O AutoML framework generates derived models by optimizing hyperparameters primarily through two distinct approaches. The first approach utilizes "expert recommendation models" (e.g., GBM_1 to GBM_5), which are initially trained based on pre-validated hyperparameter combinations (Expert Heuristics) to ensure stability. The second approach yields "grid search models" (e.g., GBM_grid), which are precisely tuned to the given dataset by executing a random grid search across a vast hyperparameter space, such as learning rates and maximum tree depths.

In this study, various machine and deep learning models were explored and trained utilizing the H2O AutoML framework. The performances of the top-ranking models, evaluated based on the validation set, are presented in Table IV.

From the experimental results, Stacked Ensemble-based models recorded the highest performance in terms of the AUCPR metric (0.9505). However, the ultimate objective of this study extends beyond the mere maximization of predictive performance; rather, it seeks to analyze the internal decision-making structure of the black-box model and rigorously verify its "causal fairness." Complex ensemble architectures possess inherent limitations, rendering clearly isolating the contributions of individual variables or tracking causal impacts under counterfactual scenarios difficult.

Therefore, this study excluded structurally uninterpretable complex models from the evaluation. Instead, a single ensemble model, the GBM, was selected as the final baseline model. The GBM maintained excellent predictive power (AUCPR = 0.9474), comparable to the top-performing ensemble, while yielding highly stable SHAP value calculations. An evaluation of this final model using a completely independent test dataset demonstrated exceptional generalization performance, achieving an AUCPR of 0.9700. All subsequent causal intervention simulations were conducted based on this selected baseline model.

Table IV. Comparison of predictive model performances

Model (Algorithm)	AUCPR	AUC	LogLoss
StackedEnsemble_AllModels	0.950463	0.954131	0.2392
StackedEnsemble_BestOfFamily	0.950177	0.953893	0.240094
GBM_3*	0.947364	0.952092	0.256482
GBM_1	0.945998	0.949	0.252743
GBM_5	0.944191	0.948823	0.26318
GBM_2	0.943961	0.948797	0.261501
GBM_4	0.943861	0.948583	0.26713
GBM_grid	0.943569	0.947951	0.26689
XGBoost_grid	0.943207	0.946617	0.258683
GLM_1	0.9432	0.947289	0.254146

4.2 Analysis of Multivariate Counterfactual Intervention Results

In this section, the effects of causal interventions aimed at flipping the student dropout predictions were verified step-by-step. First, a primary counterfactual intervention was conducted by controlling solely for the four sensitive variables related to "parents' background" (qualifications and occupations), which were considered potential sources of systemic bias. Table V presents a detailed sample of 5 randomly selected students out of a total of 22 students whose predictions are successfully flipped by adjusting only these environmental factors (Appendix A).

Table V. Cases of Prediction Flips Due to Single Environmental Factor Intervention (4 Sensitive Variables)

Student_ID	Final Prediction Outcome	Change in Dropout Probability	Detailed Sensitive Variable Modifications (Qualification & Occupation)
26	Dropout→Graduate	0.708 → 0.498	Mother's qual:37→1, Mother's occ:9→4 | Father's qual:37→1, Father's occ:9→4
29	Dropout→Graduate	0.515 → 0.277	Mother's qual:37→1, Mother's occ:9→194 | Father's qual:37→38, Father's occ:9→163
104	Dropout→Graduate	0.665 → 0.465	Mother's qual:37→1, Mother's occ:4→194 | Father's qual:19→38, Father's occ:5→163
112	Dropout→Graduate	0.580 → 0.366	Mother's qual:1→19, Mother's occ:4→3 | Father's qual:1→19, Father's occ:9→10
119	Dropout→Graduate	0.877 → 0.465	Mother's qual:37→5, Mother's occ:9→2 | Father's qual:37→5, Father's occ:9→2

The experimental results indicated that out of the entire high-risk cohort, only 22 students (approximately 9.7%) had their predictions flipped from "Dropout" to "Graduate" simply by positively altering their parental environment. This explicitly demonstrates the limitations of conventional approaches to XAI and fairness, highlighting that merely adjusting sensitive variables (i.e., parents' background) is insufficient to fully rectify entrenched algorithmic biases or effectively provide actionable recourse for high-risk students.

To overcome the limitations of interventions relying solely on environmental variables, this study additionally conducted a six-variable-based multivariate counterfactual exploration. This approach strategically combined uncontrollable fixed variables (parents' background) with actionable variables at the institutional level, specifically academic achievement (number of approved curricular units in the 2nd semester) and financial status (debtor status). By integrally adjusting academic achievement, financial status, and parental background factors, this multivariate counterfactual exploration aimed to effectively transcend the constraints of single-variable environmental interventions. Table VI presents core samples illustrating the primary patterns among the 128 cases where the simulation successfully flips the prediction outcomes.

Table VI. Cases of Prediction Flips Due to Multivariate Integrated Intervention (6 Variables: Academic + Financial + Environment)

Student_ID	Final Prediction Outcome	Change in Dropout Probability	Detailed Multivariate Perturbation (Academic + Financial + Environment)
28	Dropout→Graduate	0.953 → 0.474	[Academic] 3→6 courses (+3) | [Debt] 1→0 | Mother's qual/occ: 3→1, 3→9 | Father's qual/occ: 1→37, 6→9
11	Dropout→Graduate	0.812 → 0.354	[Academic] 4→5 courses (+1) | [Debt] 1→0 | Mother's qual/occ: 1→1, 3→9 | Father's qual/occ: 38→37, 9→9
138	Dropout→Graduate	0.967 → 0.263	[Academic] 2→4 courses (+2) | [Debt] 0→0 | Mother's qual/occ: 4→1, 2→9 | Father's qual/occ: 4→37, 10→9
139	Dropout→Graduate	0.976 → 0.301	[Academic] 1→5 courses (+4) | [Debt] 0→0 | Mother's qual/occ: 38→1, 9→9 | Father's qual/occ: 38→37, 9→9
159	Dropout→Graduate	0.824 → 0.090	[Academic] 3→4 courses (+1) | [Debt] 0→0 | Mother's qual/occ: 1→1, 1→9 | Father's qual/occ: 19→37, 6→9
201	Dropout→Graduate	0.974 → 0.366	[Academic] 2→5 courses (+3) | [Debt] 1→0 | Mother's qual/occ: 19→1, 9→9 | Father's qual/occ: 37→37, 7→9
297	Dropout→Graduate	0.984 → 0.432	[Academic] 0→ 5courses (+5) | [Debt] 0→0 | Mother's qual/occ:3→1, 90→9 | Father's qual/occ:19→37, 90→9
302	Dropout→Graduate	0.652 → 0.067	[Academic] 5→6 courses (+1) | [Debt] 1→0 | Mother's qual/occ: 38→1, 5→9 | Father's qual/occ: 1→37, 3→9
346	Dropout→Graduate	0.960 → 0.499	[Academic] 3→ 5 courses (+2) | [Debt] 1→0 | Mother's qual/occ: 37→1, 6→9 | Father's qual/occ: 37→37, 3→9
376	Dropout→Graduate	0.988 → 0.308	[Academic] 0→5 courses (+5) | [Debt] 0→0 | Mother's qual/occ: 3→1, 3→9 | Father's qual/occ: 3→37, 2→9
433	Dropout→Graduate	0.619 → 0.447	[Academic] 5→7 courses (+2) | [Debt] 0→0 | Mother's qual/occ: 1→1, 4→9 | Father's qual/occ: 1→37, 4→9
522	Dropout→Graduate	0.890 → 0.357	[Academic] 3→4 courses (+1) | [Debt] 0→0 | Mother's qual/occ: 37→1, 5→9 | Father's qual/occ: 37→37, 9→9
539	Dropout→Graduate	0.605 → 0.330	[Academic] 3→4 courses (+1) | [Debt] 0→0 | Mother's qual/occ: 19→1, 90→9 | Father's qual/occ:22→37, 90→9
541	Dropout→Graduate	0.552 → 0.455	[Academic] 6→11 courses (+5) | [Debt] 0→0 | Mother's qual/occ: 18→1, 1→9 | Father's qual/occ: 18→37, 1→9
567	Dropout→Graduate	0.987 → 0.480	[Academic] 0→5 courses (+5) | [Debt] 0→0 | Mother's qual/occ: 19→1, 90→9 | Father's qual/occ: 19→37, 90→9

The complex intervention demonstrated an overwhelming intervention success rate compared to the single intervention, successfully flipping the predictive decisions of the AI model. The three primary academic insights derived from these results are as follows:

- The law of minimal intervention: The analysis reveals that when financial deficits (e.g., debt clearance) and penalties associated with parental backgrounds are mitigated, even a minimal addition in academic achievement—such as completing one additional course (+1 Unit)—is sufficient to shift extreme dropout probabilities into the safe graduation zone in numerous observed cases. This suggests that, rather than uniformly forcing high-risk students to achieve unreasonable grade improvements, prioritizing the removal of financial debts and providing targeted mentoring tailored to individual circumstances ensures significantly higher policy efficiency and student retention rates.

- Rescuing extreme risk cases: High-risk students with initial dropout prediction probabilities exceeding 0.95, who were practically classified as "confirmed dropouts" by the model, also exhibited a significant pattern of breaking through the probability threshold (0.5) and entering the predicted graduate cohort when academic achievement improvements were combined with structural corrections of their parental backgrounds. This proves that the predictive model does not operate merely through the independent summation of variables; rather, it possesses an internal non-linear causal synergy that produces abrupt probability flips when multiple environmental and academic variables interact.
- Baseline normalization: In the counterfactual scenarios, the manipulated target values for parental backgrounds that led to successful interventions were not from the highest socioeconomic tier (e.g., professionals or Ph.D. holders) but from the most commonly observed ordinary educational and occupational groups (the mode) within the dataset. This demonstrates that a prediction flip is attainable simply by normalizing a specific student's adverse environmental deficits to a "baseline" level, without the need to elevate them to an exceptional degree. Consequently, this strongly supports the premise that eliminating the "structural penalty" imposed by AI on vulnerable groups constitutes the core of securing causal fairness.

In summary, the empirical results of the multivariate counterfactual analysis demonstrate that the AI model decision boundary is not merely a reflection of linear academic progress but deeply intertwined with socioeconomic proxy variables. By identifying the specific "Causal Flips" that combine academic effort with environmental normalization, our framework proves that algorithmic fairness can be achieved through targeted, minimal interventions rather than generic institutional policies. These findings provide a concrete evidence-based foundation for designing early warning systems that are not only accurate but also causally just and actionable.

4.3 Identification of Causal Mechanisms at the Individual Level: ΔShap and Counterfactual Contrast

To quantify the internal mechanisms through which multivariate interventions flip prediction outcomes, a focused causal audit was conducted on the case of Student 138, an extreme high-risk instance with an initial dropout probability of 0.967. Table VII and the SHAP waterfall plot in Fig. 3 illustrate the specific changes in the inference structure before and after the intervention.

Table VII. Causal Understanding through Counterfactuals (Student 138)

Feature Name	Actual State (Before)	Intervention (Counterfactual)	Causal Impact (ΔSHAP)
Curricular units 2nd sem (approved)	2 Units	4 Units	-3.00 (Flip)
Debtor	Yes (1)	No (0)	-0.40 (Reduction)
Father's qualification	37	1	-0.28
Mother's qualification	4	1	-0.2
Father's occupation	10	9	-0.17
Mother's occupation	2	9	-0.16
Outcome (Dropout Probability)	0.9671	0.2629	Δ-0.7042

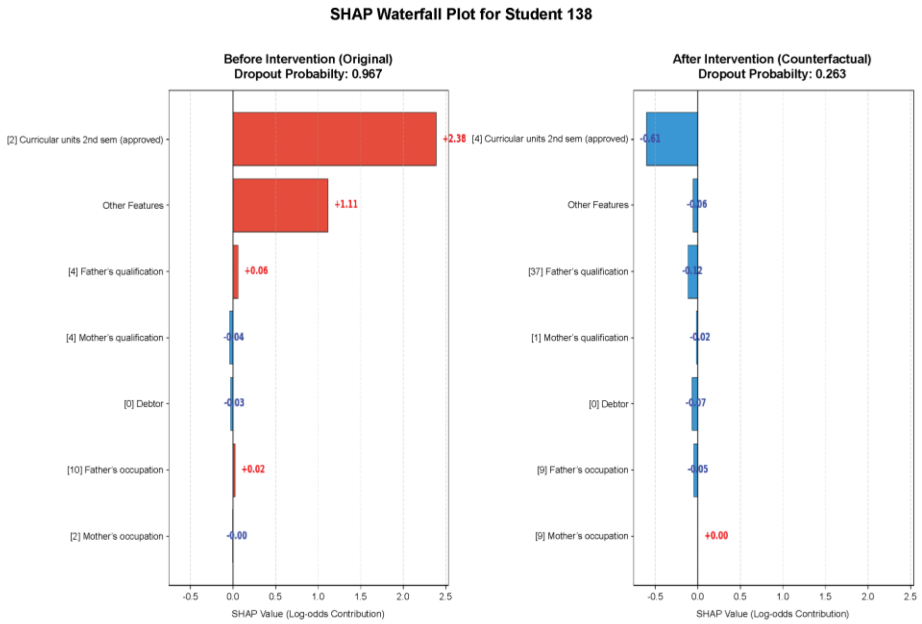

Figure 3. Student 138 waterfall plot

Dominant dropout drivers and environmental penalties: Prior to the intervention, the model assigned a substantial dropout contribution (+2.38 log-odds) to the deficiency in the "number of approved curricular units in the 2nd semester." Additionally, environmental penalties—such as the father's specific low-tier occupational group—acted as overlapping factors that exacerbated the risk of dropout.

- Decisive recourse levers and structural penalty mitigation: As the number of approved units in the 2nd semester increased from 2 to 4, the ΔSHAP value for this variable underwent a dramatic reversal to -3.00, transforming a dropout driver into a powerful catalyst for graduation. Concurrently, by normalizing parental backgrounds to the dataset mode, the factors that previously heightened risk shifted toward neutral or risk-reducing (blue) indicators, effectively "unshackling" the student from environmental constraints.
- · Derivative synergetic effects: The total contribution of the remaining variables (Other Features), which were not directly manipulated, also shifted collectively from the dropout direction toward the graduation direction. This empirically demonstrated that the multivariate intervention fundamentally reconfigured the complex nonlinear interaction structures inherent within the tree-based ensemble model.

This local-level analysis provides deep insights into how a specific student's predicted dropout is flipped through nonlinear interactions (Local Understanding). However, to confirm whether the mechanisms observed in a single instance are universal systemic patterns governing the entire predictive model, rather than accidental outliers, an expansion of the analytical scope is essential. According to Baron's [8] Causal Auditing framework, a complete interpretation of a model decision-making structure is only achieved when microscopic findings are cross-validated through a macroscopic evaluation of the entire population. Therefore, to prove that the causal synergy and bias-neutralization effects discovered at the "microscopic" level are indeed global phenomena consistently applied across the entire high-risk cohort (Global Understanding), the analysis was extended as follows.

- Primary drivers: The analysis revealed that the "number of approved curricular units in the 2nd semester (★)," which was directly intervened upon, exhibited a ΔSHAP value of -2.711, most powerfully reducing the risk of student dropout. This was followed by the "debt resolution (★)" manipulation (-0.122), which showed the second-highest contribution to risk reduction. These findings empirically demonstrate that academic improvement and financial stability constitute the core causal pathways that drive the model decision making toward the "Graduate" outcome within multivariate interventions.
- Debiasing effect: A particularly noteworthy observation is the shift in parental background variables (e.g., Father's qualification, Mother's occupation). Although the model initially imposed dropout penalties based on these variables, their contributions to dropout risk simultaneously decreased (indicated by green bars) after being normalized to the average level of successful graduates through the intervention. This signified that the intervention did not merely improve academic records but effectively "unshackled" the student from the environmental disadvantages embedded within the model.
- Spillover effect: Furthermore, the SHAP values of variables that were not directly intervened upon (e.g., Tuition fees up to date, Curricular units 1st sem) also underwent a chain reaction of changes. This confirms the presence of complex interactions between input variables—a characteristic of tree-based ensemble models

like GBM—and proves that the calibration of specific features triggers a non-linear synergy across the entire predictive space. The aggregate causal impact of these multivariate interventions on the model's decision logic is quantitatively summarized in Figure 4.

The ΔSHAP analysis in this study holds significant academic value as it reverse-engineers the rationale behind why the dropout prediction AI "gives up" on certain students. Moreover, it provides concrete quantitative evidence (e.g., the -2.711 contribution to risk reduction) to identify which institutional policies, specifically combining debt resolution with targeted mentoring, must be integrated to overcome these risks.

In conclusion, this case-specific causal audit transcends the limitations of static explanations by visualizing the dynamic reconfiguration of the model logic. By quantifying how structural biases can be neutralized through actionable interventions, we establish a rigorous evidentiary bridge between black-box predictions and equitable educational support.

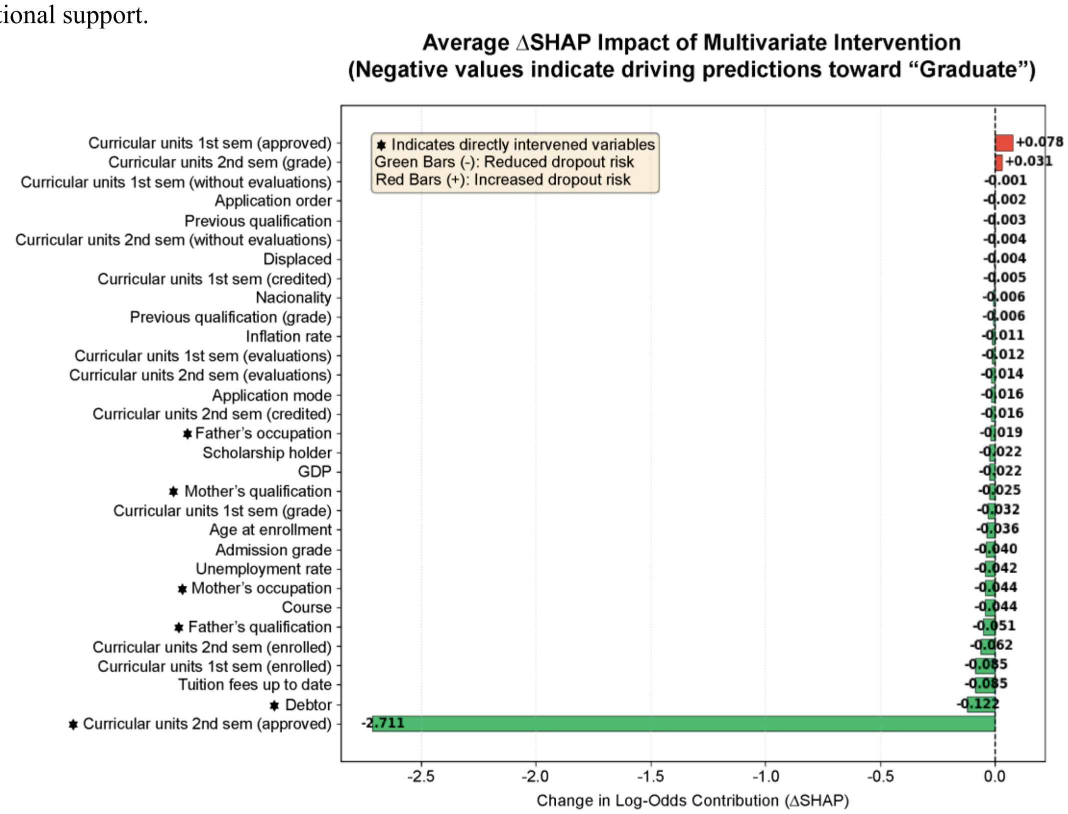

Figure 4. Student 138 waterfall plot

4.4 Global Causal Generalization and Evaluation Metric Analysis

According to Baron's [8] Causal Auditing framework, a comprehensive model interpretation is achieved only when microscopic analysis of individual cases is integrated with a macroscopic evaluation of the entire population. Although Section IV.C provides a local understanding of a specific student (Student 138), this section verifies the global understanding of how the baseline predictive model causally perceives the entire cohort of high-risk students.

First, to confirm whether the effects of multivariate interventions appeared consistently across the entire group, the intervention trajectories of all target students were visualized using subplots of individual conditional expectation (ICE) and PDP. In Fig. 5, the gray lines (ICE) represent the changes in predicted probability for individual students, while the thick orange line (PDP) represents the average causal effect for the entire group.

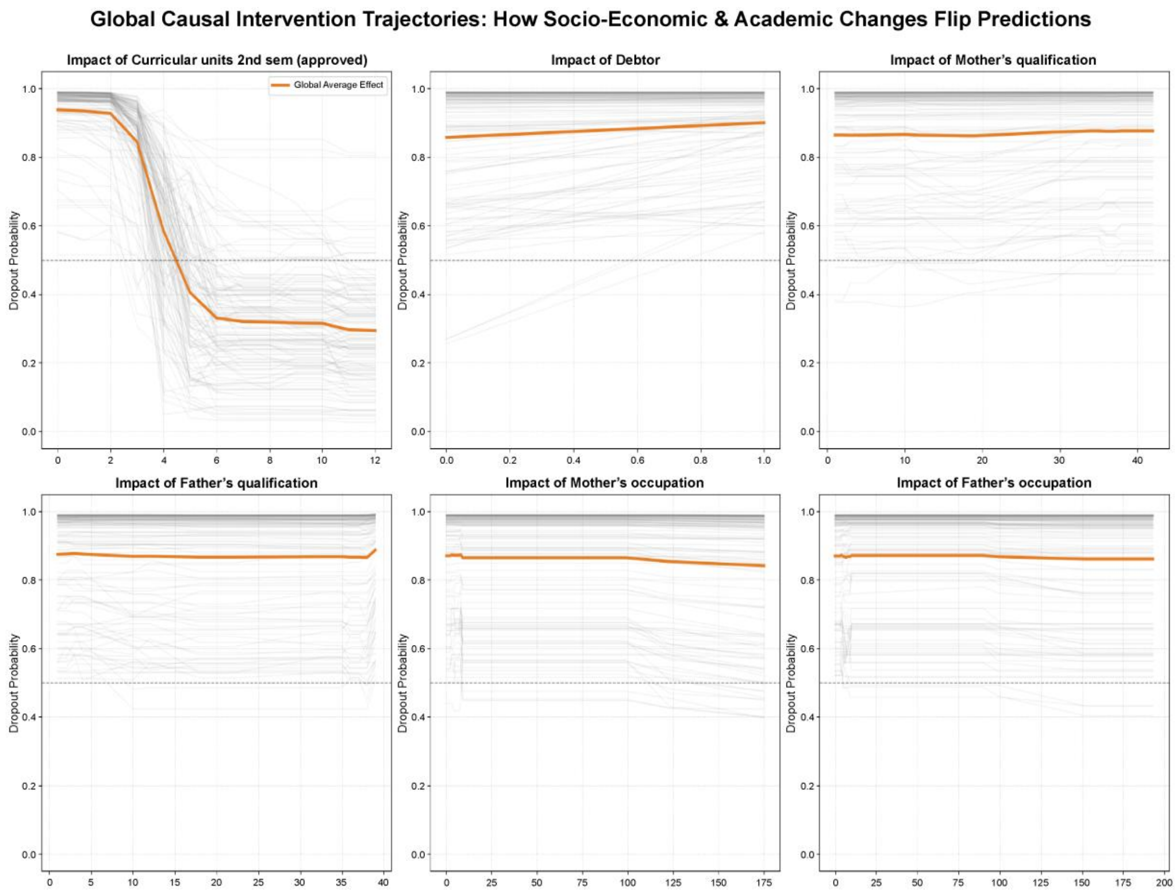

Figure 5. Global causal intervention trajectories: how socio-economic & academic changes flip predictions

This visualization demonstrates three key patterns that prove the generalizability of this study:

- Global patterns and optimal intervention point (Sweet Spot): In the chart for the "number of approved curricular units in the 2nd semester," a non-linear "cliff" phenomenon is observed where the dropout probability drops sharply below the threshold (0.5) starting from the point where the number of units exceeds 3 to 4. This presents a universal causal relationship applicable to all high-risk students and suggests the optimal "sweet spot" that university administrative academic encouragement policies must target.
- Global proof of proxy bias: In the charts for socioeconomic variables, such as parental occupation, the average probability for the entire group decreases by approximately 15% when variable values shift toward specific categories. This globally proves that an entire class of students is receiving systematic and discriminatory penalties within the learning model.

To quantitatively verify the global patterns identified visually, Table VIII presents the aggregate effectiveness results of applying multivariate causal interventions to the entire high-risk target group (N = 128), based on the evaluation metrics defined in Section III.D.

Table VIII. Causal Understanding through Counterfactuals (Student 138)

Variable Type	**Feature Name**	**Actual State (Before)**	**Intervention (Counterfactual)**	**Causal Impact (Δ SHAP)**
Academic	Curricular units 2nd sem (approved)	2 Units	4 Units	-3.00 (Flip)
Financial	Debtor	No (0)	No (0)	-0.4 (Reduction)
Socio-economic	Mother's qualification	4	1	-0.2
Socio-economic	Father's qualification	4	37	-0.28

Socio-economic	Mother's occupation	2	9	-0.16
Socio-economic	Father's occupation	10	9	-0.17
Outcome	Dropout Probability	0.9671	0.2629	Δ -0.7042

As shown in Table VIII, the causal intervention data for Student 138 were extracted with high precision. In particular, the dropout probability decreased from 0.9671 to 0.2629 (a reduction of -0.7042), resulting in a complete prediction flip. This result strongly supports the core hypothesis of this study through the following mechanisms:

- Causal Mechanism Identification: Table 8 demonstrates that the root cause of Student 138 being classified as a "Dropout" was not merely due to low values in individual variables, but was rooted in the model's internal causal logic. Specifically, adjusting academic achievement (Curricular units 2nd sem) from 2 units to 4 units resulted in a SHAP value shift of -3.00, inducing a decisive prediction flip. This serves as concrete evidence of how "Understanding via Intervention," as emphasized by Baron [8], operates within a real-world educational data environment.
- Neutralizing Socio-economic Bias: An interesting observation is the shift in socioeconomic variables, such as parental background (Qualification, Occupation) and debt status (Debtor). Although the individual SHAP changes (ΔSHAP) for these variables were relatively smaller (-0.16 to -0.40) compared to academic achievement, their combined effect created a cumulative contribution reduction of approximately -1.21. This suggests that, beyond an individual student's personal efforts (academic), there is sufficient causal influence to transition high-risk students to the predicted graduate group simply by removing the "environmental shackles" imposed by the system.
- Ensuring "Causal Transparency" of the Black-Box Model: These analytical results prove that complex non-linear models like GBM do not rely solely on correlations; rather, they logically modify their decision-making rationale in response to changes in specific variables. The "Causal Impact (ΔSHAP)" presented in Table VIII quantifies how the model perceives each variable as a "Reason," enabling us to simultaneously demonstrate the legitimacy and rectifiability of AI predictions to the educators.

Table VIII provided a local understanding of "why a specific student's (Student 138) prediction was flipped." We further verified the global understanding of "how the baseline model causally perceives the entire high-risk cohort" through Table IX.

Table IX. Aggregate Effectiveness Of Causal Interventions (N=128)

Intervention Domain	Primary Variables	Avg. Risk Reduction	Causal Significance
Academic Performance	Approved Units (2nd Sem)	0.379	High (Structural)
Socio-economic Proxies	Parents' BG, Debtor Status	0.147	Moderate (Proxy Neutralization)
Combined (Full Intervention)	All 6 Variables	0.527	Critical (Flip Target)

Based on the analysis of all 128 high-risk students, the multivariate intervention proposed in this study achieved a 100% FSR and a robust ΔP of 0.527. When categorized by domain, academic achievement improvement accounted for the largest share (-0.379), proving that academic support remains a powerful structural solution for student retention. However, the most notable finding is that a statistically significant risk reduction of -0.147 occurred through the adjustment of socioeconomic proxy variables alone. This indicates that the AI predictive model was adopting biased weightings based on debt status and parental background in addition to the student’s actual capabilities. Through this proxy neutralization—the correction of such causal errors—this study globally demonstrates that a fairer decision boundary can be established. The combination of visual trajectory analysis (Fig. 5) and quantitative evaluation (Table IX demonstrates the generalizability of the proposed framework, which can causally audit and correct latent biases within black-box models.

5. Conclusion

This study proposed a causal XAI-based auditing framework to overcome the structural limitations of correlation-based explanations in machine learning-based student dropout prediction models and identify and rectify the socioeconomic proxy biases latent within these models. By applying a hierarchical multivariate counterfactual intervention to the top-performing baseline model (LightGBM) built through H2O AutoML, the following key findings were derived.

First, although conventional XAI (SHAP) exhibited structural limitations by assigning high importance to uncontrollable variables based on simple correlations, we empirically demonstrated that the decisive causal triggers for flipping prediction outcomes were a combination of "actionable 2nd-semester curricular units" and "financial status (debtor status)".

Second, when only sensitive variables (environmental factors) such as parents' education and occupation were adjusted, a mere 9.7% (22 students) of the high-risk cohort was retained. However, an overwhelming synergistic effect was demonstrated when environmental corrections were combined with substantive academic and financial support (multivariate intervention), resulting in prediction flips of 100% (all 128 students) toward graduation.

Third, by neutralizing socioeconomic proxy variables—such as debt status and parental background—to the average levels of the entire population, the dropout probability for all students decreased significantly (average risk reduction $\Delta P = -0.147$). This globally proves the existence of latent structural bias within the black-box model.

The study findings provide significant academic and practical contributions to the fields of data science and higher education administration. First, although conventional XAI research primarily focuses on "Explanation," this study expands the research horizon to "Causal Auditing". By being the first to empirically apply Baron's [8] causal understanding framework to the education domain, this study verifies "how to flip" a decision. In particular, it presents a new methodological framework for evaluating the fairness of AI models by mathematically quantifying the "Causal Mismatch" between correlation-based importance and actual causal contributions. Second, this study provides specific practical guidelines for dropout prevention policies in universities. According to the simulation results, forcing high-risk students with student loan issues to achieve unreasonable grade improvements across all subjects is inefficient. Instead, the "Law of Minimal Intervention," which prioritizes resolving financial issues (debt resolution) and combining it with minimal, unit-targeted mentoring (e.g., +1 Unit), demonstrated the highest retention rate. This suggests that university administrations must establish a multi-dimensional hybrid support network by integrating data from student counseling centers (academic) and scholarship departments (financial). Furthermore, these results can serve as a critical practical foundation for designing "bidirectional human–AI collaboration" systems, where data-driven algorithmic judgments are combined with interventions by educational experts when establishing future AI-based personalized learning support policies [19], [20], [21]. Lastly, the multivariate counterfactual intervention and proxy bias auditing framework proposed in this study can be flexibly extended beyond the education domain to FinTech and financial sectors, specifically for credit scoring and loan

Although this study successfully demonstrates the potential of causal XAI for bias detection and prediction flipping, several limitations exist that present opportunities for future research. First, the experiments were conducted using a dataset from a specific higher education institution in Portugal. Cross-institutional validation across different countries or heterogeneous educational environments is required. Future research should verify the universality of the causal trajectories by applying this framework to datasets encompassing diverse regional and cultural contexts. Second, although this study established parental background and financial status as the primary proxy variables, the analytical scope could be expanded in future studies through causal graph modeling. This would incorporate more dynamic and microscopic behavioral data, such as students' psychological states, attendance patterns, and learning management system (LMS) access logs. Third, to establish the methodological generalizability of the proposed causal auditing framework, further research extending its application to other industrial domains is necessary. In particular, follow-up studies should utilize real-world credit scoring data from the aforementioned FinTech and financial sectors to quantitatively diagnose systemic biases inherent in financial AI models and verify algorithmic accountability. Finally, follow-up empirical studies based on randomized controlled trials (RCTs) or A/B testing must be conducted. These studies are essential to measure the actual changes in dropout rates when the generated multivariate counterfactual scenarios are practically implemented as institutional policies in real-world university settings.

APPENDIX A

Data Dictionary for Socio-Economic Variables (Full Specification)

The full mapping specifications between the original numeric codes and categorical labels for the socioeconomic sensitive attributes (parents' qualifications and occupations) in the Portuguese higher education institution dataset utilized in this study are as follows. The original Data Dictionary is included in its entirety to ensure the transparency of data preprocessing and reproducibility of the experiments.

Table A.I . Full Mapping of Parents' Qualification Codes (34 Categories)

Code	Educational Level (Short Label)	Original UCI Description
1	High School Graduate	Secondary Education - 12th Year of Schooling or Eq.

2	Bachelor's Degree	Higher Education - Bachelor's Degree
3	Bachelor's Degree (Other)	Higher Education - Degree
4	Master's Degree	Higher Education - Master's
5	Doctorate	Higher Education - Doctorate
6	Attending Higher Education	Frequency of Higher Education
9	12th Year Drop-out	12th Year of Schooling - Not Completed
10	11th Year Drop-out	11th Year of Schooling - Not Completed
11	7th Year (Old System)	7th Year (Old)
12	11th Year (Other)	Other - 11th Year of Schooling
13	2nd Year Complementary High School	2nd year complementary high school course
14	10th Year Completed	10th Year of Schooling
18	General Commerce Course	General commerce course
19	Middle School Graduate (9/10/11th Year)	Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv.
20	Complementary High School Course	Complementary High School Course
22	Technical-Professional Course	Technical-professional course
26	7th Year Completed	7th year of schooling
27	2nd Cycle General High School	2nd cycle of the general high school course
29	9th Year Drop-out	9th Year of Schooling - Not Completed
30	8th Year Completed	8th year of schooling
31	General Course of Administration and Commerce	General Course of Administration and Commerce
33	Supplementary Accounting and Administration	Supplementary Accounting and Administration
34	Unknown	Unknown
35	Illiterate (Cannot read or write)	Cannot read or write
36	Less than 4 years (Can read)	Can read without having a 4th year of schooling
37	Elementary School Graduate (4/5th Year)	Basic education 1st cycle (4th/5th year) or equiv.
38	Basic Education (6/7/8th Year)	Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv.
39	Technological Specialization Course	Technological specialization course
40	Higher Education Degree (1st Cycle)	Higher education - degree (1st cycle)
41	Specialized Higher Studies Course	Specialized higher studies course
42	Professional Higher Technical Course	Professional higher technical course
43	Master's (2nd Cycle)	Higher Education - Master (2nd cycle)
44	Doctorate (3rd Cycle)	Higher Education - Doctorate (3rd cycle)

TABLE A.II. Full Mapping of Parents' Occupation Codes (Mapped According To Portuguese National Classification Of Occupations And Isco-08)

Code	Occupational Group (Short Label)	Original UCI Description
0	Student	Student
1	Senior Officials and Managers	Representatives of Legislative Power and Executive Bodies, Directors
2	Intellectual and Scientific Specialists	Specialists in Intellectual and Scientific Activities
3	Intermediate Level Technicians and Professionals	Intermediate Level Technicians and Professions
4	Administrative Support Staff	Administrative staff

5	Service/Security/Sales Workers	Personal Services, Security and Safety Workers and Sellers
6	Skilled Agricultural/Forestry/Fishery Workers	Farmers and Skilled Workers in Agriculture, Fisheries and Forestry
7	Skilled Workers in Industry/Construction/Crafts	Skilled Workers in Industry, Construction and Craftsmen
8	Machine/Assembly Operators	Installation and Machine Operators and Assembly Workers
9	Unskilled Workers	Unskilled Workers
10	Armed Forces	Armed Forces Professions
90	Other Situation	Other Situation
99	Blank/No Response	(blank)
122	Health Professionals	Health professionals
123	Teachers and Educators	Teachers
125	ICT Specialists	Specialists in information and communication technologies (ICT)
131	Intermediate Science/Engineering Technicians	Intermediate level science and engineering technicians
132	Intermediate Health Technicians	Technicians and professionals, of intermediate level of health
134	Intermediate Legal/Social/Cultural Technicians	Intermediate level technicians from legal, social, sports, cultural
141	General Office and Data Processing Clerks	Office workers, secretaries in general and data processing operators
143	Data/Accounting/Financial/Statistical Clerks	Data, accounting, statistical, financial services and registry-related operators
144	Other Administrative Support Staff	Other administrative support staff
151	Personal Service Workers	Personal service workers
152	Sellers	Sellers
153	Personal Care Workers	Personal care workers and the like
171	Skilled Construction Workers	Skilled construction workers and the like, except electricians
172	Skilled Metalworkers	Skilled workers in metallurgy, metalworking and similar
174	Skilled Electrical/Electronic Workers	Skilled workers in electricity and electronics
175	Skilled Manufacturing Workers (Food/Wood/Clothing)	Workers in food processing, woodworking, clothing and other industries
191	Cleaning Workers	Cleaning workers
192	Unskilled Agricultural/Fishery Workers	Unskilled workers in agriculture, animal production, fisheries and forestry
193	Unskilled Manufacturing/Construction/Transport Workers	Unskilled workers in extractive industry, construction, manufacturing
194	Meal Preparation Assistants	Meal preparation assistants

References

1. E. Tiukhova, P. Vemuri, N. López Flores, A. S. Islind, M. Óskarsdóttir, S. Poelmans, B. Baesens, and M. Snoeck, "Explainable Learning Analytics: Assessing the stability of student success prediction models by means of explainable AI," *Decision Support Systems*, vol. 182, Art. no. 114229, 2024. DOI: 10.1016/j.dss.2024.114229.
2. J. G. C. Krüger, A. de Souza Britto Jr., and J. P. Barddal, "An explainable machine learning approach for student dropout prediction," *Expert Systems with Applications*, vol. 233, Art. no. 120933, 2023. DOI: 10.1016/j.eswa.2023.120933.
3. A. M. Rabelo and L. E. Zárate, "A Model for Predicting Dropout of Higher Education Students," *Data Science and Management*, vol. 8, no. 1, pp. 72–85, 2025.
4. A. Turkmenbayev, E. Abdykerimova, S. Nurgozhayev, G. Karabassova, and D. Baigozhanova, "The application of machine learning in predicting student performance in university engineering programs: A rapid review," *Frontiers in Education*, vol. 10, Art. no. 1562586, 2025. DOI: 10.3389/feduc.2025.1562586.
5. M. Nagy and R. Molontay, "Interpretable dropout prediction: Towards XAI-based personalized intervention," *International Journal of Artificial Intelligence in Education*, vol. 34, no. 2, pp. 274–300, 2024. DOI: 10.1007/s40593-023-00331-8.
6. M. M. Islam, F. H. Sojib, M. F. H. Mihad, M. Hasan, and M. Rahman, "The integration of explainable AI in educational data mining for student academic performance prediction and support system," *Telematics and Informatics Reports*, vol. 18, Art. no. 100203, 2025. DOI: 10.1016/j.teler.2025.100203.
7. Y. Kim and S. Woo, "On use of eXplainable artificial intelligence to explain student dropout," *Journal of Korean Association for Artificial Intelligence Education*, vol. 5, no. 2, pp. 12–25, 2024. DOI: 10.52618/aied.2024.5.2.2.
8. S. Baron, "Explainable AI and causal understanding: Counterfactual approaches considered," *Minds and Machines*, vol. 33, no. 2, pp. 347–377, 2023. DOI: 10.1007/s11023-023-09637-x.
9. G. Cornacchia, V. W. Anelli, E. Di Sciascio, F. M. Donini, and T. Di Noia, "Auditing fairness under unawareness through counterfactual reasoning," *Information Processing & Management*, vol. 60, no. 2, Art. no. 103224, 2023. DOI: 10.1016/j.ipm.2022.103224.
10. R. Paul, D. Singh, and A. K. Bairagi, "Analyzing dropout of students and an explainable prediction of academic performance utilizing artificial intelligence techniques," *Frontiers in Education*, vol. 10, 2025. DOI: 10.3389/feduc.2025.1698505.
11. A. Bettahi, F.-Z. Belouadha, and H. Harroud, "A modular and explainable machine learning pipeline for student dropout prediction in higher education," *Algorithms*, vol. 18, no. 10, Art. no. 662, 2025. DOI: 10.3390/a18100662.
12. Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," *Information Fusion*, vol. 81, pp. 59–83, 2022. DOI: 10.1016/j.inffus.2021.11.003.
13. L. E. J. Bynum, J. R. Loftus, and J. Stoyanovich, "A new paradigm for counterfactual reasoning in fairness and recourse," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI 2024)*, 2024.
14. N. Pham, A. Nguyen-Duc, and X.-B. D. Le, "Fairness for machine learning software in education: A systematic mapping study," *Journal of Systems and Software*, vol. 219, Art. no. 112244, 2025. DOI: 10.1016/j.jss.2024.112244.
15. S. Gunasekara and M. Saarela, "Explainable AI in education: Techniques and qualitative assessment," *Applied Sciences*, vol. 15, no. 3, Art. no. 1239, 2025. DOI: 10.3390/app15031239.
16. K. Kesgin, S. Kiraz, S. Kosunalp, and B. Stoycheva, "Beyond performance: Explaining and ensuring fairness in student academic performance prediction with machine learning," *Applied Sciences*, vol. 15, no. 15, Art. no. 8409, 2025. DOI: 10.3390/app15158409.
17. G. Raftopoulos, G. Davrazos, and S. Kotsiantis, "Fair and transparent student admission prediction using machine learning models," *Algorithms*, vol. 17, no. 12, Art. no. 572, 2024. DOI: 10.3390/a17120572.
18. V. Realinho, J. Machado, L. Baptista, and M. V. Martins, *Predict Students' Dropout and Academic Success*. UCI Machine Learning Repository, 2022.
19. R. Parkavi, P. Karthikeyan, and A. Sheik Abdullah, "Enhancing personalized learning with explainable AI: A chaotic particle swarm optimization based decision support system," *Applied Soft Computing*, vol. 156, Art. no. 111451, 2024. DOI: 10.1016/j.asoc.2024.111451.
20. R. Zong, Y. Zhang, L. Shang, F. Stinar, N. Bosch, and D. Wang, "Bidirectional human–AI collaboration for equitable student performance prediction via deep uncertainty learning," in *Proceedings of IJCAI 2025*. California: International Joint Conferences on Artificial Intelligence Organization, 2025, pp. 10026–10034. DOI: 10.24963/ijcai.2025/1114.
21. J. Tsao, "Trajectories of AI Policy in Higher Education: Interpretations, Discourses, and Enactments of Students and Teachers," *Computers and Education: Artificial Intelligence*, vol. 9, Art. no. 100496, 2025
22. Ovcharenko, N., Samoilenko, O., Moskva, O., & Chebotarenko, O. (2020). Innovative technologies in vocal training: Technological culture formation of future musical art teachers. Journal of History Culture and Art Research, 9(3), 115-126.
23. Özdek, A. (2015). The role of folk music as cultural heritage in the curriculum of vocational high schools of music: Pattern of Azerbaijan-Turkey. Anadolu Journal of Educational Sciences International, 5(3), 84-112.
24. Sun, L. (2021). A probe into the integration of traditional music culture in vocal music teaching in colleges and

universities. Region-Educational Research and Reviews, 3(2), 65.
25. Kalyuzhnaya, V. P. (2025). “Folk Music Culture” as an Academic Discipline in the System of Secondary Vocational Education: Problems and Perspectives. Russian Musicology, (3), 148-156.
26. Zhang, P. (2020, November). The Significance and Ways of Constructing the Northwest Traditional Music Curriculum System in Lanzhou Vocational Colleges. In International Conference on Education Studies: Experience and Innovation (ICESEI 2020) (pp. 582-586). Atlantis Press.
27. Li, K. (2024). Innovative Strategies in Teaching Vocal Music for Higher Vocational Music Performance Majors. Art and Performance Letters, 5(2), 146-152.