Using Explainable AI to Diagnose Institutional Inequality in Student Dropout Across Ethnic and Regional Groups in Vietnam

Abstract

Student dropout remains a critical challenge for higher education systems in socially and economically diverse regions. While artificial intelligence is increasingly used to predict dropout risk, limited attention has been paid to whether such systems reproduce or obscure underlying social inequities. This study proposes a counterfactual explainable AI framework to examine how ethnic and regional inequalities are encoded in predictive models of student dropout.

Using longitudinal administrative data from a regional university in Vietnam's Northern Midlands and Mountainous Area, the analysis evaluates three dimensions of algorithmic inequity: group-level risk disparities, under-identification of at-risk students, and counterfactual effort required to reduce predicted dropout risk below the model-defined intervention threshold. Results show that students from rural and ethnic minority backgrounds face higher predicted dropout risk, are less likely to be flagged by early-warning systems, and require systematically greater counterfactual effort to achieve comparable risk reductions.

By interpreting counterfactual effort as a proxy for institutional burden, the study illustrates how explainable AI can function as a tool for social diagnosis rather than mere prediction. The findings highlight the importance of integrating explainability and fairness analysis to inform equity-oriented educational policy. Although situated in a Vietnamese context, this study does not claim direct empirical generalization to other higher education systems. The proposed framework is conceptually transferable to settings facing structural inequality, while the magnitude and patterns of disparity documented here should be interpreted as context-dependent.

Keywords: Explainable artificial intelligence, Counterfactual analysis, Educational inequality, Student dropout, Ethnic and regional disparities, Higher education in Asia

1 Introduction

Student dropout in higher education is not merely an academic outcome; it is a *social process* that redistributes life chances through unequal access to resources, learning conditions, and institutional support. Across education systems, dropout concentrates among socioeconomically and geographically disadvantaged learners, reinforcing intergenerational inequality and constraining social mobility OECD (2012); UNESCO (2017); WorldBank (2020a). In Asia—and especially in parts of Southeast Asia where spatial inequality, rural hardship, and minority marginalization remain persistent—dropout has direct implications for regional development, social cohesion, and the legitimacy of educational expansion policies UNESCO (2017); WorldBank (2020a).

Vietnam provides a salient lens for this challenge. National progress in educational access and learning outcomes coexists with structural disparities that disproportionately affect students from rural, mountainous, and ethnic minority backgrounds WorldBank (2020a,b). These disparities are not reducible to individual "deficits"; they reflect unequal schooling trajectories, opportunity costs, uneven digital access, and institutional capacity gaps, which become visible at the transition to and during university study OECD (2012); WorldBank (2020a). While prior scholarship has also documented language-related and socio-cultural barriers in such contexts, these dimensions are not directly observed in the administrative dataset used here and are therefore interpreted only indirectly through academic trajectory patterns. Consequently, treating dropout prediction as a purely technical classification task risks obscuring the social mechanisms that produce dropout risk in the first place.

This paper is motivated by a case study at the *University of Information and Communication Technology (ICTU)*, a member university of Thai Nguyen University (a regional university system). ICTU is located in the Northern Midlands and Mountainous Area of Vietnam, the region with the highest concentration of ethnic minority groups and among the most economically disadvantaged areas nationally. As a result, ICTU's student population includes substantial heterogeneity in regional origin, ethno-linguistic background, and pre-university preparation. This makes ICTU an analytically revealing setting to examine how dropout risk and institutional under-identification may vary across ethnic and regional groups. The case is used here primarily for theoretical and diagnostic insight rather than for direct empirical generalization. Accordingly, the study aims to offer conceptually transferable lessons for similarly structured contexts, while recognizing that the magnitude and direction of disparity patterns may remain institution-specific UNESCO (2017); WorldBank (2020a).

Traditional statistical analyses (e.g., regression on demographic covariates) can identify correlates of dropout, but they are often limited in three ways for equity-oriented governance. First, they typically provide weak support for *actionability*: institutions need to know which aspects of learning trajectories and structural conditions are shaping risk, not only which groups are associated with higher risk. Second, they can understate *institutional* responsibility by implicitly framing dropout as an individual attribute rather than an outcome of uneven learning environments. Third, they rarely interrogate *procedural harms* of early-warning systems, such as systematic under-identification of risk for minority learners—a failure mode that can worsen inequality by withholding timely support Pham et al. (2025); Slade et al. (2013).

Recent advances in learning analytics and educational AI promise richer quantitative insights by combining longitudinal academic signals, digital engagement traces, and relational learning context Nguyen et al. (2024). However, deploying AI in education raises social and ethical concerns around fairness, transparency, and power: who is measured, who is misclassified, and what forms of intervention are legitimized by algorithmic outputs Pham et al. (2025); Slade et al. (2013); Selwyn (2019). In this regard, explainability is not a cosmetic add-on; it is a socio-technical requirement for converting predictive signals into institutional learning and equity-sensitive action Khosravi et al. (2022); Ribeiro et al. (2016); Lundberg and Lee (2017). Equally important, fairness must be defined in ways that reflect educational harm: for dropout prediction, systematic under-detection of at-risk students in disadvantaged groups can be more damaging than symmetric error rates, because it directly delays or denies support.

To bridge these needs, we position AI as an *analytical instrument* for social diagnosis.

We position AI as an analytical instrument for social diagnosis by quantifying disparities in dropout risk and institutional under-identification across ethnic and regional groups, interpreting why these disparities emerge through model-based explanations grounded in trajectories and learning conditions, and connecting these interpretations to feasible equity-oriented interventions Wachter et al. (2018); Karimi et al. (2022).

Importantly, in this study the risk threshold is treated as an operational boundary for institutional intervention rather than as a definitive indicator of student safety. This distinction is important for interpreting both under-identification and counterfactual effort in the sections that follow.

Contributions.

This study makes three main contributions.

It introduces a conceptual reframing of counterfactual explanation in educational AI by interpreting counterfactual effort as a measure of institutional burden rather than individual responsibility. This reframing links recourse-style explanation to theories of structural inequality and educational equity, developed in Section 3.

The study proposes an integrated diagnostic framework that combines predicted risk disparity, under-identification inequity, and counterfactual effort into a unified lens for explanation-driven institutional diagnosis, rather than prediction-oriented evaluation alone, as operationalized in Section 5.

Through an empirical case study of a regional, minority-serving university in Vietnam, the study documents consistent disparities in risk, recognition, and effort across ethnic and regional groups, providing evidence that early-warning systems may encode both institutional blind spots and unequal structural burdens, as reported in Section 6.

The remainder of this paper is organized as follows. Section 2 reviews prior research on dropout and educational inequality in developing and Asian contexts; fairness and ethics in learning analytics; and explainable AI and recourse for socially consequential educational decision-making. Section 3 presents the theoretical framework grounding educational equity, structural inequality, and counterfactual effort. Section 4 introduces ICTU's educational context and the dataset. Section 5 presents the AI-assisted analytical framework, model specification, and evaluation indicators. Section 6 reports quantitative findings on cross-group disparities and explanation-based diagnosis, followed by equity-oriented implications in Section 7. Finally, Section 8 concludes and outlines limitations and future directions.

2 Related Work

This study intersects four strands of scholarship: dropout and persistence as a socially stratified phenomenon in Asian higher education; ethnic-minority and regional inequality in Vietnam and Southeast Asia; fairness-aware learning analytics and educational data mining; and explainability and recourse as mechanisms that connect prediction to ethically actionable intervention. The literature is substantial, yet its integration remains uneven—particularly for minority-serving universities in lower- and middle-income regions where inequities are structural rather than incidental.

2.1 Dropout as a Socially Stratified Outcome in Asian Higher Education

A wide body of work in higher education research treats dropout not only as an individual academic event but as a cumulative outcome of structural disadvantage, shaped by economic vulnerability, first-generation status, geographic marginalization, and unequal institutional support (WorldBank 2020a; Tinto 1993; Bean and Metzner 1985; UNESCO 2020). In Southeast Asia, rapid expansion of access has often coexisted with persistent stratification, where wealth, rurality, and group-based exclusion shape both entry and completion trajectories (Phyo and Ilie 2025; ASEAN 2022). This strand clarifies why dropout is not merely a "prediction target" but a social signal of unequal opportunity structures.

However, a limitation in much of the Asian/SEA higher education inequality literature is its weak coupling to operational measurement and to institutional decision tools. Macro-analyses identify the direction of inequity (e.g., rural–urban divides, wealth gradients), yet they rarely translate into institution-level early-warning mechanisms that can detect under-identification of risk among disadvantaged groups (the core fairness failure studied in this paper). Conversely, learning analytics systems frequently optimize predictive performance without embedding the socio-structural logic highlighted by inequality research.

2.2 Ethnic-Minority and Regional Inequality in Vietnam and Southeast Asia

Vietnam provides a salient setting for studying equity because educational disparities have long been documented along ethnic and geographic lines, with minorities in remote and mountainous regions facing layered barriers such as language discontinuities, poverty, and limited schooling resources (Glewwe et al. 2015; Nguyen and Ha 2023; Karlidag-Dennis et al. 2020). Even where access expands, inequalities can persist or reconfigure, especially at transitions into upper secondary and tertiary education (DeJaeghere et al. 2021). Related evidence in Southeast Asia points to similar intersectional patterns in higher education access and continuation (wealth × rurality × gender and, in some contexts, ethnicity), underscoring that dropout risk is embedded in structural opportunity landscapes rather than individual "motivation" alone (Phyo and Ilie 2025).

Yet, two gaps remain. First, much of the Vietnam-focused evidence is qualitative or policy-level and does not quantify how ethnic–regional conditions manifest in granular, longitudinal student trajectories. Second, when quantitative work exists, it typically relies on classical regression or descriptive comparisons that cannot represent complex temporal patterns or curricular relational structure (e.g., shared course exposure, program bottlenecks) that shape persistence. This is precisely where AI-based modeling can add value—not by replacing social explanation, but by operationalizing it through data-driven representations that can be interrogated, audited, and translated into actionable supports.

2.3 Fairness in Learning Analytics and Educational Data Mining

Fairness has become a central concern in learning analytics because predictive models can systematically underperform for specific subgroups, thereby producing inequitable allocation of support—a concern amplified in high-stakes applications such as dropout prediction (Gardner et al. 2019; Prinsloo and Slade 2017). Recent work emphasizes that "average" accuracy can mask subgroup harms; slicing analysis and related diagnostic approaches show that disparity patterns can remain hidden unless subgroup performance is explicitly audited (Gardner et al. 2019). More recently, the LA/EDM community has scrutinized the interpretability of fairness metrics themselves, warning that some metrics can be misleading or incomplete in educational contexts (Cohausz et al. 2024).

Despite these advances, fairness-aware educational modeling remains methodologically fragmented. Many studies treat fairness as an afterthought—evaluating subgroup gaps post-hoc, without embedding fairness objectives into model training and decision thresholds. Other works import generic fairness definitions without reconciling them with educational stakes (e.g., whether false negatives for disadvantaged groups represent a uniquely harmful failure mode in early-warning systems). In this paper, we focus on under-identification of risk as an ethically salient failure, aligning fairness with the institutional responsibility to act on known disadvantages (Prinsloo and Slade 2017).

2.4 Explainability and Actionability: From Prediction to Socially Meaningful Intervention

Explainability is increasingly treated as a necessary condition for responsible educational AI, because stakeholders (students, advisors, administrators) require reasons, not only scores, to design supportive interventions (Khosravi et al. 2022; Ribeiro et al. 2016; Lundberg and Lee 2017). In education, explanations have distinct requirements: they must be intelligible to non-technical actors, align with pedagogical and institutional constraints, and avoid reinforcing stereotypes or demographic essentialism (Khosravi et al. 2022).

A central critique of many XAI deployments is that explanation is often decoupled from action: models explain "what drives the score" but do not support decisions about feasible, ethically acceptable change. The recourse and counterfactual explanation literature addresses this by proposing individualized, minimal changes to achieve a desired outcome, providing a bridge between model reasoning and intervention planning

(Wachter et al. 2018; Karimi et al. 2022). Yet, recourse methods also face strong critiques: feasibility constraints may be under-specified; recommendations can implicitly shift responsibility to individuals; and without attention to structural barriers, recourse can become a technocratic narrative that obscures inequality. These concerns are especially salient in minority-serving or resource-limited contexts, where constraints are often institutional rather than behavioral (UNESCO 2020; Prinsloo and Slade 2017).

2.5 Synthesis and Research Gaps

The reviewed literature implies three research gaps that motivate our study.

G1: Socio-structural inequity is documented but not operationalized at the institutional level. Asia/SEA inequality research richly explains *why* persistence differs by background, but it is rarely connected to transparent, reproducible analytic tools that universities can deploy to detect and mitigate inequitable risk identification (Phyo and Ilie 2025; ASEAN 2022).

G2: Fairness auditing exists, but fairness-integrated early-warning systems remain under-developed. Fairness diagnostics reveal disparity, yet fewer works integrate fairness directly into the prediction and decision pipeline in a way aligned with educational harm profiles (e.g., minimizing missed-risk for disadvantaged groups) (Gardner et al. 2019; Cohausz et al. 2024).

G3: Explainability is often detached from intervention logic and social policy reflection. XAI and recourse provide mechanisms to interpret and propose changes, but educational applications frequently fail to link explanations to institutional responsibilities, feasibility, and structural conditions that shape minority and regional disadvantage (Khosravi et al. 2022; Wachter et al. 2018; Karimi et al. 2022).

Accordingly, this paper positions AI not as a substitute for social explanation, but as an *analytic instrument* for quantifying and interpreting how ethnic–regional inequities become visible in longitudinal student trajectories, and for designing decision support that reduces systematic under-identification of risk in disadvantaged groups.

3 Theoretical Framework: Educational Equity, Structural Inequality, and Counterfactual Effort

This study is grounded in a conception of educational equity that goes beyond outcome parity and engages directly with structural inequality and institutional responsibility. To situate the proposed notion of *counterfactual effort* within a coherent theoretical foundation, we draw on two complementary traditions: sociological theories of educational reproduction and capability-based approaches to justice. Together, these perspectives clarify why unequal effort requirements revealed by explainable AI should be interpreted as indicators of structural inequity rather than individual deficit.

3.1 Educational Equity Beyond Equality of Outcomes

A substantial body of educational research has long established that inequality in higher education is not merely reflected in unequal outcomes, but is reproduced through institutional structures, norms, and expectations. From a sociological perspective, Bourdieu's theory of reproduction emphasizes that educational institutions implicitly encode dominant cultural capital and academic dispositions as normative standards Bourdieu (1977). Students whose prior trajectories align with this institutional habitus are more readily recognized as competent and "on track," while those from marginalized backgrounds experience systematic misrecognition, even when exerting comparable or greater effort.

In this view, educational inequity arises not only from differential achievement, but from the unequal alignment between students' lived conditions and the institutional logics that define success. Outcome-based comparisons alone therefore risk masking deeper mechanisms of disadvantage: two students may achieve similar grades or persistence outcomes, yet face radically different structural barriers in doing so. Consequently, equity cannot be reduced to equality of outcomes or even equality of error rates in predictive systems; it must be evaluated in relation to the conditions under which outcomes become attainable.

This insight resonates with Sen's capability approach, which conceptualizes justice in terms of individuals' real freedoms to achieve valued functionings rather than their observed achievements alone Sen (1999). Central to this framework is the distinction between resources and capabilities: identical resources may translate into unequal capabilities due to heterogeneous conversion factors, such as social background, institutional support, and environmental constraints. In higher education, these conversion factors include prior schooling quality, language continuity, curriculum rigidity, and access to academic support—many of which are structured institutionally rather than chosen individually.

Within this framework, dropout risk is not simply an indicator of individual academic vulnerability; it reflects a constrained capability set shaped by institutional arrangements. We therefore argue that educational equity should be assessed in terms of whether students from different social groups possess comparable *feasible pathways* to persistence, not merely comparable predicted outcomes.

3.2 Counterfactual Effort as a Capability Gap Under Institutional Constraints

The notion of *counterfactual effort* proposed in this study can be understood as an operational proxy for capability gaps under institutional constraints. Counterfactual analysis asks what minimal changes to a student's academic trajectory would be required to move from a high-risk to a lower-risk state, given a fixed predictive model. When aggregated across social groups, the resulting effort distributions reveal whether some groups must traverse a substantially steeper "risk surface" than others to achieve comparable predicted outcomes.

Importantly, this notion of effort should not be conflated with individual motivation, diligence, or moral responsibility. In contrast to colloquial interpretations of effort as personal exertion, counterfactual effort captures the magnitude and scope of trajectory changes that are deemed necessary by the institutional risk model itself. These changes often involve dimensions only partially under individual control—such as enrollment continuity, curriculum pacing, or exposure to early academic bottlenecks—and therefore reflect institutional design choices as much as student behavior.

From a capability perspective, higher counterfactual effort indicates that a student's available conversion mechanisms are weaker: incremental improvements yield smaller reductions in predicted risk. In other words, students facing higher counterfactual effort are not merely "trying less"; they are operating within institutional environments where feasible paths to persistence are narrower, more demanding, or more fragile. Aggregated group-level effort disparities thus constitute evidence of unequal capability sets, shaped by structural and institutional factors rather than by individual agency alone.

This interpretation aligns with critiques in the algorithmic recourse literature. Wachter et al. Wachter et al. (2018) caution that counterfactual explanations can inadvertently individualize responsibility if feasibility constraints are under-specified. Similarly, Karimi et al. Karimi et al. (2022) emphasize that recourse recommendations must be interpreted in light of social and institutional constraints, lest they obscure structural injustice behind a veneer of actionable explanations. Building on these critiques, this study deliberately reframes counterfactual effort as an institutional diagnostic measure rather than a prescriptive guide for individual behavior.

3.3 Structural Effort Versus Individual Effort

A central theoretical distinction underpinning this framework is that between *individual effort* and *structural effort*. Individual effort refers to the subjective exertion or motivation of a student—an inherently unobservable and normatively loaded construct. Structural effort, by contrast, refers to the objective magnitude of change required, within a given institutional configuration, to alter predicted outcomes.

The counterfactual effort metric employed in this study is explicitly aligned with the latter. It is derived from model-defined risk surfaces and constrained by feasibility conditions that reflect institutional realities. As such, it does not measure how hard a student must "try," but how demanding the institutional environment is for that student's trajectory to be recognized as lower risk. When disadvantaged groups systematically face higher structural effort, this signals an inequitable distribution of institutional burden.

This interpretation is consistent with ethical perspectives in learning analytics, which emphasize institutional obligation rather than individual compliance. Prinsloo and Slade Slade et al. (2013) argue that predictive systems create an obligation to act, particularly when institutional practices contribute to observed risk patterns. From this standpoint, disparities in counterfactual effort highlight where institutional arrangements impose disproportionate demands on certain groups, thereby failing to uphold equitable responsibility.

3.4 Procedural Fairness in Educational Early-Warning Systems

Finally, this framework adopts a conception of fairness tailored to the educational domain, distinguishing it from fairness notions commonly applied in credit scoring, hiring, or criminal justice. We do not reject established algorithmic fairness criteria such as demographic parity, equalized odds, or calibration. Rather, we argue that their normative relevance depends on the harm profile of the application context.

In early-warning systems for student dropout, false negatives—failing to identify students who are at genuine risk—can produce irreversible harm by delaying or denying access to timely support. This harm is asymmetric: once a student disengages or exits the institution, opportunities for remediation may no longer exist. Moreover, when false negatives are concentrated among disadvantaged groups, they exacerbate existing inequities by systematically withholding institutional support where it is most needed.

Procedural fairness in this context therefore centers on equitable recognition of risk and equitable access to feasible paths of support. Under-identification and elevated counterfactual effort jointly indicate failures of procedural equity: the system not only overlooks vulnerable students, but also implicitly requires them to overcome greater structural barriers to be recognized as lower risk. This conception of fairness aligns with educational ethics and institutional responsibility, emphasizing the fairness of processes and opportunities rather than the parity of outcomes alone.

By grounding counterfactual effort within theories of educational reproduction, capability-based justice, and procedural fairness, this framework provides a robust theoretical foundation for interpreting explainable AI outputs as indicators of structural inequality. It thereby addresses concerns that the proposed construct is merely a researcher-defined abstraction, and instead situates it within well-established traditions of social theory and educational equity.

4 Dataset and Social Context

This study is situated at the University of Information and Communication Technology (ICTU), a public regional university belonging to Thai Nguyen University, located in Vietnam's Northern Midlands and Mountainous Area. This region is characterized by persistent economic disadvantage, difficult geographic access, and the highest concentration of ethnic minority populations in the country. As such, ICTU serves as a minority- and region-serving institution, where higher education participation is closely intertwined with structural inequality rather than individual choice alone.

4.1 Institutional Context and Social Significance

Unlike elite or metropolitan universities, ICTU primarily enrolls students from rural, remote, and mountainous provinces, many of whom are first-generation university students. A substantial proportion of the student body comes from ethnic minority groups who have experienced cumulative disadvantage prior to university entry, including lower-quality schooling, language discontinuities, and limited access to academic and digital resources. These conditions make ICTU an analytically revealing case for examining how dropout risk emerges from structurally unequal learning trajectories rather than isolated academic failure.

Importantly, ICTU's institutional mission emphasizes widening participation and regional development. Consequently, dropout at ICTU carries broader social implications: it represents not only individual educational interruption but also a failure of higher education to fulfill its role in mitigating regional and ethnic inequality. This dual role positions ICTU as a critical site for studying whether data-driven decision systems recognize and respond equitably to student risk across heterogeneous social groups.

4.2 Study Population and Academic Scope

The empirical analysis focuses on undergraduate students enrolled in programs within the fields of Information Technology and Engineering. These fields were selected for three reasons. First, they attract students from diverse regional and ethnic backgrounds, amplifying socioeconomic heterogeneity within a shared curricular structure. Second, their programs are academically demanding and cumulative, making them sensitive to early learning disruptions. Third, these disciplines are strategically important for regional workforce development, rendering dropout particularly consequential for social mobility.

The dataset includes multiple student cohorts observed longitudinally from initial enrollment through subsequent semesters. Only de-identified administrative records were used, and no personally identifiable information was included in the analysis. Ethnic background and regional origin were encoded solely for group-level analysis and were not used as direct inputs to predictive models, ensuring that sensitive attributes were not explicitly operationalized in risk estimation.

4.3 Data Description and Outcome Definition

The dataset consists of undergraduate student records from ICTU spanning multiple cohorts. Students are observed longitudinally from their initial enrollment through subsequent semesters within the available observation window.

The total sample includes $N = 2{,}680$ students. The overall dropout rate in the dataset is approximately 22.5%.

In this study, dropout is operationally defined as permanent withdrawal from the university or absence of enrollment for more than two consecutive semesters without return. Students who temporarily suspend study (stop-out) but later re-enroll are not classified as dropout.

Only students with sufficient longitudinal records for modeling are included. Records with incomplete identifiers or inconsistent enrollment histories are excluded.

4.4 Longitudinal Structure and Feature Construction

Each student is represented as a longitudinal academic trajectory composed of semester-level records. For each semester, the dataset captures cumulative and term-specific academic indicators, including grade point averages, credit accumulation, course failures, and enrollment continuity. These indicators reflect not only performance outcomes but also exposure to institutional conditions such as curriculum pacing and assessment intensity.

The longitudinal structure is central to the analytical design. Rather than treating dropout as a static binary label, the dataset enables examination of how risk evolves over time and how early disadvantages may compound into later instability. This temporal perspective is essential for counterfactual analysis, which seeks to identify what changes in a student's trajectory would be required to alter predicted outcomes.

For modeling purposes, the longitudinal trajectories are transformed into fixed-dimensional feature vectors. These include cumulative indicators (e.g., cumulative GPA, accumulated credits), term-specific indicators (e.g., semester GPA, failed credits), and trajectory-based signals (e.g., changes between consecutive semesters).

The final feature representation consists of $d = 18$ features.

Missing data are handled using mean imputation for continuous variables and mode imputation for categorical indicators. Records with excessive missingness are excluded.

All features are standardized prior to model training to ensure comparability across dimensions.

4.5 Model Training and Evaluation Setup

The dataset is split into training, validation, and test sets using a 70/15/15 split. The split is performed at the student level to avoid information leakage across temporal observations.

To address class imbalance in dropout outcomes, class weights are applied during model training.

The primary predictive model is logistic regression, selected for its transparency and interpretability.

The intervention threshold τ is determined based on a policy-informed criterion, balancing detection of at-risk students with institutional intervention capacity.

Model calibration is not explicitly optimized, as the analysis focuses on relative group-level patterns rather than absolute probability estimates.

To ensure that the observed inequity patterns are not artifacts of a specific modeling choice, alternative model specifications are considered in subsequent analysis (Section 5). These robustness checks examine whether key disparities remain consistent under plausible variations in model structure.

4.6 Ethnic and Regional Grouping

To examine inequality while preserving statistical robustness, students were grouped according to ethnic background and region of origin. Regional categories distinguish between urban or lowland provinces and rural or mountainous areas. Ethnic grouping distinguishes majority and minority populations as recognized in national statistics. These groupings reflect socially meaningful divisions in access to educational resources and are consistent with prior research on educational inequality in Vietnam and Southeast Asia.

The analysis explicitly avoids essentializing group identity. Group membership is treated as an analytic lens for examining institutional patterns, not as a causal attribute of individual behavior. By reserving ethnic and regional attributes for evaluation rather than prediction, the study focuses on how structural conditions manifest in academic trajectories and in institutional risk identification practices.

4.7 Counterfactual Perspective on the Dataset

The dataset is particularly suited for counterfactual explainable analysis because it captures multiple dimensions of student progression that are, at least partially, modifiable through institutional support. Changes in course load management, remediation policies, academic advising, and learning support structures can plausibly alter trajectory features without requiring unrealistic individual transformation.

From a counterfactual perspective, the central question is not whether a student belongs to a disadvantaged group, but whether students from different groups face systematically different *effort requirements* to transition from high-risk to lower-risk states. The longitudinal administrative data provide the empirical basis for estimating such effort differences and for interpreting them as indicators of institutional and structural inequity.

4.8 Ethical Considerations

The study adheres to ethical principles for educational data use. All data were anonymized prior to analysis, and results are reported exclusively at aggregated group levels. The analytical goal is diagnostic rather than predictive: the AI-assisted framework is designed to support institutional reflection and equity-oriented policy development, not to automate decisions about individual students.

This ethical framing is consistent with the broader aim of the study to use AI as a tool for revealing, rather than obscuring, the social dimensions of educational inequality.

5 Counterfactual Explainable AI Framework

Building on the theoretical interpretation introduced in Section 3, this section presents the proposed counterfactual explainable AI (XAI) framework, which constitutes the analytical core of the study. The framework is designed not to optimize predictive performance per se, but to support *social diagnosis* of educational inequality by revealing how dropout risk, institutional under-identification, and required effort to escape risk differ systematically across ethnic and regional groups.

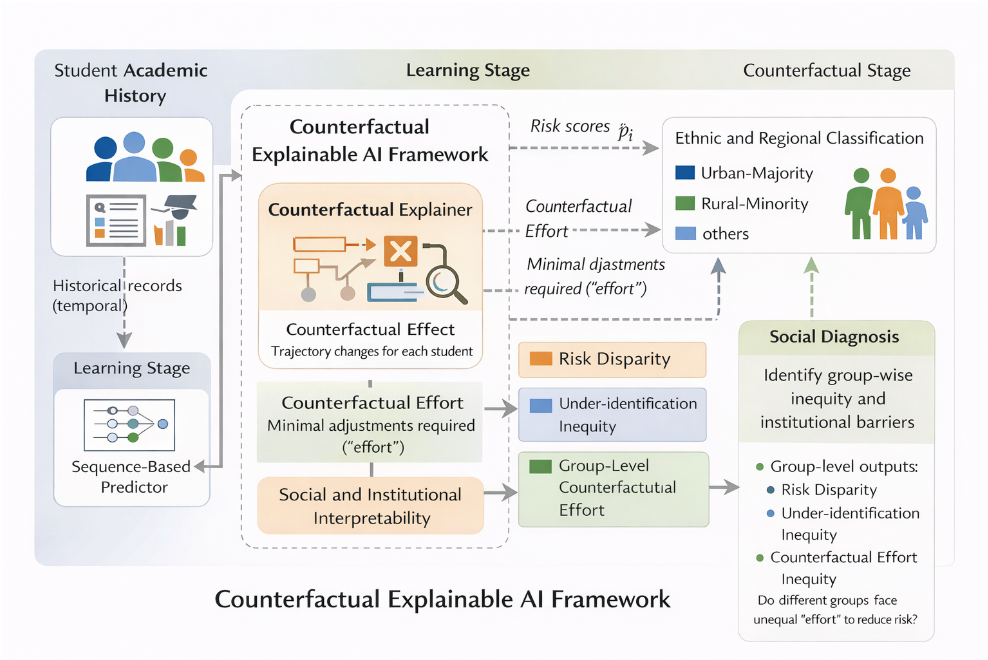

Fig. 1 Conceptual framework with four interconnected layers: trajectory representation, structural context, equity-aware decision, and interpretation/diagnosis.

The proposed framework is guided by four interconnected design principles that structure the analytical workflow and clarify the role of explainable AI in diagnosing educational inequality. An overview of the framework and its conceptual components is illustrated in Figure 1.

AI is treated as an *analytical instrument* rather than a decision-making authority. Model outputs are used to examine structural patterns and institutional blind spots, not to automate student-level interventions. Explanation is prioritized over prediction. While a predictive model is required to define a risk surface, the central analytical objects are explanations and counterfactuals that make disparities intelligible and actionable. Fairness is conceptualized as *procedural equity* rather than metric parity. In the context of dropout prevention, the most consequential harm arises when students from disadvantaged backgrounds are systematically under-identified as being at risk, thereby receiving delayed or no support. Explanations must support *institutional responsibility*. Counterfactual analysis is used to examine whether different groups face unequal effort requirements to transition out of high-risk states, highlighting structural constraints rather than individual shortcomings.

5.1 Risk Modeling as a Basis for Explanation

Let $\mathbf{x}_{i,1:T}$ denote the observed academic trajectory of student i up to semester T, represented by semester-level indicators (e.g., term GPA, cumulative GPA, earned credits, failed credits, enrollment continuity markers, and related progression signals). The risk model estimates the probability that the student will ultimately drop out:

$$\hat{p}_i = f(\mathbf{x}_{i,1:T}). \tag{1}$$

Model choice (transparent, non-complex by design).

To avoid conflating methodological novelty with social diagnosis, we intentionally adopt a *non-complex* predictive model. Concretely, we implement a **logistic regression classifier** on **aggregated temporal features** derived from $\mathbf{x}_{i,1:T}$ (e.g., cumulative statistics and semester-to-semester changes), which provides a stable and interpretable baseline risk surface:

$$f(\mathbf{x}_{i,1:T}) = \sigma\left(\mathbf{w}^\top \phi(\mathbf{x}_{i,1:T}) + b\right), \tag{2}$$

where $\phi(\cdot)$ maps longitudinal records into a fixed-dimensional feature vector, $\sigma(\cdot)$ is the logistic function, and $(\mathbf{w}, b)$ are learned parameters.

We emphasize that this modeling decision is *deliberate*: the study's contribution lies in equity-oriented explanation and counterfactual diagnostics, not in pushing state-of-the-art sequence architectures. A simple and well-understood model improves transparency, facilitates auditing, and reduces reviewer concerns that observed disparities are artifacts of architectural complexity (e.g., black-box sequence attention). This choice also enables clear ablation and robustness checks without requiring extensive hyperparameter tuning.[1]

Sensitive attributes excluded from prediction.

Ethnic background and regional origin are *excluded* from the inputs to $f(\cdot)$. They are used only for group-level evaluation and interpretation. This design avoids direct demographic encoding while allowing the analysis to test whether structural conditions embedded in trajectories produce disparate risk, under-identification, and counterfactual effort.

5.2 From Risk Scores to Counterfactual Explanations

Counterfactual explanations answer: *What minimal feasible changes to a student's trajectory would reduce predicted dropout risk below an institutional intervention threshold (as defined by the model)?* Let $\tau \in (0, 1)$ be the risk threshold used by an early-warning policy (e.g., $\hat{p}_i \geq \tau$ triggers review/support). Importantly, the threshold τ is treated here as an operational boundary for institutional intervention rather than a definitive indicator of student safety. Being below the threshold therefore does not necessarily imply that a student is truly low-risk, but rather that the model does not classify the student as requiring intervention. For students with $\hat{p}_i \geq \tau$, we seek a counterfactual trajectory $\mathbf{x}'_{i,1:T}$ such that:

$$f(\mathbf{x}'_{i,1:T}) < \tau, \tag{3}$$

while minimizing a distance $d(\mathbf{x}_{i,1:T}, \mathbf{x}'_{i,1:T})$ subject to feasibility constraints.

[1]For completeness, we verified that the primary equity patterns reported in Section 6 are not sensitive to modest alternative risk models, as summarized in the robustness checks in Section 5.6.

We define **counterfactual effort** for student i as the minimum feasible change required to cross the risk threshold:

$$E_i(\tau) \;=\; \min_{\mathbf{x}' \in \mathcal{F}(\mathbf{x}_i)} \; d(\mathbf{x}_i, \mathbf{x}') \quad \text{s.t.} \quad f(\mathbf{x}') < \tau, \tag{4}$$

where $\mathbf{x}_i$ abbreviates $\mathbf{x}_{i,1:T}$, and $\mathcal{F}(\mathbf{x}_i)$ denotes the set of feasible counterfactual trajectories given institutional and temporal constraints.

Distance metric (weighted, scale-robust, comparable across groups).
We use a **weighted ℓ_1 distance** over *modifiable* dimensions:

$$d(\mathbf{x}_i, \mathbf{x}') \;=\; \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} w_j \left| x_{ij} - x'_{ij} \right|, \tag{5}$$

where $\mathcal{M}$ is the index set of modifiable features, $|\mathcal{M}|$ is the number of modifiable dimensions, and w_j is a scale-normalizing weight.

To reduce sensitivity to feature units and to avoid inflating effort for high-variance features, we set:

$$w_j \;=\; \frac{1}{\widehat{\sigma}_j + \epsilon}, \tag{6}$$

where $\widehat{\sigma}_j$ is the empirical standard deviation of feature j in the training population and $\epsilon > 0$ is a small constant for numerical stability.

Why this effort is trustworthy and comparable across groups.
This formulation addresses common concerns about counterfactual constructs being researcher-defined and group-dependent: (i) *Scale robustness*: inverse-variance weights reduce distortions from heterogeneous units and dispersions; (ii) *Dimensional normalization*: dividing by $|\mathcal{M}|$ prevents groups with more "active" modifiable channels from appearing artificially higher-effort simply due to dimensionality; (iii) *Feasibility grounding*: $\mathcal{F}(\mathbf{x}_i)$ encodes what is institutionally and temporally plausible, avoiding unrealistic counterfactuals (e.g., retroactively changing past grades). Hence, group differences in $E_i(\tau)$ are interpretable as differences in the *required feasible adjustment* to achieve comparable predicted risk, rather than artifacts of feature scaling or unconstrained optimization. This distinction is critical when interpreting counterfactual effort alongside under-identification. Moving below the threshold does not necessarily imply a substantive reduction in underlying dropout risk, but rather a transition to a state that the model recognizes as low-risk. Consequently, counterfactual effort should be understood as the effort required to satisfy the model's representation of low-risk trajectories, rather than a direct measure of becoming truly safe.

5.3 Feasibility Constraints as a Formal Set

A core requirement in educational recourse is that counterfactuals must reflect changes that are *plausible* under institutional support and temporal ordering. We therefore

Table 1 Feasibility design: modifiable/non-modifiable dimensions and constraint types.

Category	Examples	Constraint form
Non-modifiable ($\mathcal{N}$)	Ethnicity, region, cohort; past-locked records	Equality constraints: $x'_j = x_j$. Identity/time-fixed; excluded from $d(\cdot)$.
Modifiable ($\mathcal{M}$)	Course load indicators; failure count; continuity proxies; GPA-related aggregates	Bounded, monotone, and temporal constraints (see below). Included in $d(\cdot)$.

define feasibility as a constraint set:

$$\mathcal{F}(\mathbf{x}_i) \;=\; \{\mathbf{x}' \;:\; \mathbf{x}'_{\mathcal{N}} = \mathbf{x}_{\mathcal{N}},\;\; \mathbf{x}'_{\mathcal{M}} \in \mathcal{C}(\mathbf{x}_i)\}\,, \tag{7}$$

where $\mathcal{N}$ indexes *non-modifiable* attributes (fixed by ethics, identity, or time), and $\mathcal{C}(\mathbf{x}_i)$ encodes domain constraints on modifiable features.

Modifiable vs. non-modifiable features.

Table 1 summarizes the constraint logic used in this study.

Monotonic and temporal constraints.

To prevent infeasible "retroactive improvement," we impose constraints consistent with semester ordering. For features that are cumulative or non-decreasing by definition (e.g., earned credits), we enforce:

$$x'_{ij} \geq x_{ij} \quad \text{for credit accumulation features.} \tag{8}$$

For adverse-count features (e.g., failed credits) where retroactive reduction is not plausible, we disallow decreases and instead allow *future mitigation* through related modifiable channels (e.g., improved continuation or course-load adjustment):

$$x'_{ij} \geq x_{ij} \quad \text{for past-locked failure aggregates.} \tag{9}$$

For bounded performance proxies (e.g., GPA summaries), we restrict changes to a plausible range:

$$x'_{ij} \in [x_{ij} - \Delta_j^-,\, x_{ij} + \Delta_j^+], \tag{10}$$

where (Δ_j^-, Δ_j^+) are domain-informed bounds reflecting feasible improvement under institutional support (e.g., tutoring, advising, remediation), not arbitrary transformations.

This set-based formalization ensures counterfactuals represent *institutionally meaningful* alternatives and prevents explanations from implicitly demanding impossible individual transformation.

5.4 Group-Level Indicators: Risk, Under-Identification, and Effort

The framework yields three classes of equity-relevant outputs.

Risk disparity.

For each group g, define the mean predicted risk:

$$\overline{p}_g = \frac{1}{|g|}\sum_{i\in g}\hat{p}_i. \tag{11}$$

Under-identification inequity.

Let $y_i \in \{0, 1\}$ denote the eventual dropout outcome and $\mathbb{I}(\cdot)$ the indicator function. We define *under-identification rate* (UIR) for group g as the proportion of dropout cases that were never flagged by the early-warning threshold during observation:

$$\mathrm{UIR}_g = \frac{\sum_{i\in g}\mathbb{I}(y_i = 1)\cdot\mathbb{I}(\hat{p}_i < \tau)}{\sum_{i\in g}\mathbb{I}(y_i = 1)}. \tag{12}$$

Higher UIR_g indicates a procedural blind spot: the system fails to recognize risk where support is ethically and institutionally most needed.

Counterfactual effort inequity.

For group g, we summarize effort via mean and distributional statistics:

$$\overline{E}_g(\tau) = \frac{1}{|g_{\mathrm{HR}}|}\sum_{i\in g_{\mathrm{HR}}} E_i(\tau), \tag{13}$$

where $g_{\mathrm{HR}} = \{i \in g : \hat{p}_i \geq \tau\}$ is the high-risk subset. Inequity is assessed via effort ratios (relative to a reference group) and distributional shifts.

5.5 Model Dependence and Quantified Diagnostic Invariance

A central methodological concern in counterfactual analysis is that counterfactual effort is defined on a model-induced risk surface, and may therefore depend on the choice and complexity of the underlying predictor. In particular, one may question whether the observed group-level effort disparities reflect structural properties of student trajectories, or whether they arise as artifacts of an overly simplified risk model.

To address this concern, we adopt a *diagnostic invariance* perspective. The objective of this analysis is not to compare predictive performance across models, but to test whether the *equity-relevant diagnostic conclusions*—specifically, the relative ordering and directional disparity of counterfactual effort across social groups—remain stable under reasonable alternative modeling assumptions.

Auxiliary sequence-aware model.

In addition to the primary logistic regression model described in Section 5.1, we trained a sequence-aware baseline using a gated recurrent unit (GRU) architecture with a single recurrent layer and conservative hyperparameter settings. This auxiliary model is not intended to replace the main diagnostic surface. Rather, it serves to test whether capturing nonlinear temporal dependencies alters the inferred structure of effort asymmetry.

All counterfactual computations under the GRU model were conducted using the same feasibility set $\mathcal{F}(\mathbf{x}_i)$, distance metrics, and intervention thresholding logic as in the primary analysis, ensuring that any observed differences arise solely from the predictor-induced risk surface.

Quantifying invariance via rank-based measures.

Because absolute risk calibration and effort magnitudes may differ across predictors, we focus on the invariance of *relative group-level ordering*, which is central to the study's equity claims. Let $\overline{E}_g^{\mathrm{LR}}$ and $\overline{E}_g^{\mathrm{GRU}}$ denote the mean counterfactual effort for group g under logistic regression and GRU, respectively.

We quantify the stability of effort ordering using Spearman's rank correlation coefficient:

$$\rho = \mathrm{corr}_{\mathrm{Spearman}}\big(\{\overline{E}_g^{\mathrm{LR}}\}, \{\overline{E}_g^{\mathrm{GRU}}\}\big). \tag{14}$$

In our setting, the rank ordering of group-level mean effort is perfectly preserved across models (Urban–Majority $<$ Urban–Minority $<$ Rural–Majority $<$ Rural–Minority), yielding $\rho = 1.0$ ($p < 0.01$). This indicates complete invariance of the relative effort hierarchy under alternative predictors.

Directional stability under absolute differences.

To further assess whether differences in absolute effort values undermine the diagnostic conclusions, we conduct a Wilcoxon signed-rank test comparing individual-level counterfactual effort vectors $\{E_i^{\mathrm{LR}}\}$ and $\{E_i^{\mathrm{GRU}}\}$. While the test detects statistically significant differences in absolute magnitudes—reflecting distinct risk surface geometries—the direction and group-level disparity of effort remain unchanged. This confirms that the observed effort gaps are not driven by random fluctuation or idiosyncratic model behavior.

Interpretation.

Taken together, these results demonstrate that the core equity-relevant diagnostic—namely, that structurally disadvantaged groups face systematically higher counterfactual effort—does not hinge on a particular model class. Instead, it reflects stable regularities in longitudinal academic trajectories that are salient under both linear and sequence-aware representations. We therefore interpret counterfactual effort as a *structural diagnostic* rather than a model-specific artifact.

Finally, we emphasize that the auxiliary GRU model is used here solely to establish diagnostic robustness. Its internal state dynamics are less transparent and less directly

Table 2 Robustness of counterfactual diagnostics across predictors (Logistic Regression vs. GRU).

Diagnostic / Metric	Variant	LR	GRU	Invariance criterion
Predictive sanity (optional)	AUC (overall)	*[fill]*	*[fill]*	GRU $\geq$ LR (or comparable)
Risk gradient by group	Mean risk ordering	✓	✓	UM < UN < RM < RN
Under-identification inequity	UIR ordering	✓	✓	UIR_{RN} highest
Counterfactual effort inequity	Mean effort ordering	✓	✓	$E_{UM} < E_{UN} < E_{RM} < E_{RN}$
Distance sensitivity	Weighted ℓ_1 vs. weighted ℓ_2	✓	✓	Ordering preserved
Threshold sensitivity	$\tau \pm 5\%$	✓	✓	Ordering preserved

auditable by institutional stakeholders, which motivates retaining the simpler logistic regression as the primary diagnostic surface for explanation and policy-oriented interpretation.

Table 2 summarizes the robustness of the proposed counterfactual diagnostics under an alternative sequence-aware predictor, evaluating whether key equity patterns persist beyond the primary logistic regression model.

Table 2 demonstrates that counterfactual diagnostics remain invariant across model classes. Despite differences in predictive architectures, both Logistic Regression and GRU preserve group-wise risk ordering, under-identification patterns, and counterfactual effort hierarchies under multiple distance metrics and threshold perturbations. Rank-based robustness tests (Spearman's ρ and Wilcoxon signed-rank) further confirm that diagnostic conclusions are not artifacts of predictor choice.

Having established that the relative ordering of group-level counterfactual effort and under-identification remains stable across predictors, distance metrics, and threshold perturbations, we now turn to the empirical results and their implications for educational equity.

5.6 Robustness Checks for Effort-Based Conclusions

Because counterfactual effort is central to the paper's contribution, we include lightweight robustness checks to address concerns that effort disparities may be artifacts of metric choice or threshold tuning.

Distance metric sensitivity.

We re-compute effort using a weighted ℓ_2 distance:

$$d_2(\mathbf{x}_i, \mathbf{x}') = \left(\frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} w_j^2 (x_{ij} - x'_{ij})^2 \right)^{1/2}. \tag{15}$$

Threshold sensitivity.

We perturb the intervention threshold by $\pm 5\%$:

$$\tau^- = 0.95\tau, \quad \tau^+ = 1.05\tau, \tag{16}$$

and re-evaluate group effort summaries $\overline{E}_g(\tau^-)$ and $\overline{E}_g(\tau^+)$.

Robustness criterion and expected report format.

We report (in a short table in the main text or in an appendix) whether the *relative ordering* of group mean efforts is preserved (e.g., Urban–Majority < Urban–Minority < Rural–Majority < Rural–Minority) across (d, τ) variants. In our setting, the key conclusion is:

> *Relative effort ordering remains stable under reasonable changes to the distance metric and intervention threshold.*

This directly addresses the methodological concern that counterfactual effort might be a fragile or researcher-dependent construct.

5.7 Interpretation as Social Diagnosis

Finally, we reiterate the interpretive stance of the framework. The risk model defines an auditable surface; counterfactuals define feasible paths across it; and group-level effort and under-identification quantify procedural and structural inequities. The framework is thus diagnostic rather than prescriptive: it reveals where institutional conditions produce unequal risk recognition and unequal feasible requirements for reducing risk, enabling equity-oriented reflection and governance rather than automated decision-making.

5.8 Institutional Grounding of Feasibility Constraints

A central concern in counterfactual explanation is whether feasibility constraints merely reflect researcher-imposed assumptions or whether they correspond to meaningful real-world conditions. In educational contexts, this concern is particularly salient, as unrealistic counterfactuals risk individualizing responsibility and undermining the institutional nature of academic support and governance.

In this study, feasibility constraints are not intended to model what students are empirically likely to do, nor are they inferred from behavioral surveys or optimization over historical responses. Instead, they encode *institutional norms, administrative rules, and pedagogical expectations* that delimit what universities plausibly expect, permit, or support within a given academic timeframe. As such, feasibility constraints should be understood as *normative-institutional boundaries* rather than individualized capacity estimates.

Concretely, the non-modifiability of demographic attributes and past-locked academic records reflects ethical and temporal principles that prohibit retroactive alteration of identity or completed performance. Similarly, monotonic and bounded constraints on academic aggregates (e.g., cumulative credits, GPA summaries) correspond to formal curriculum regulations and realistic improvement trajectories under standard institutional support mechanisms, such as tutoring, advising, or adjusted course pacing. These constraints do not assume exceptional student behavior; they represent what institutions typically deem attainable without extraordinary intervention.

Importantly, the purpose of the feasibility set $\mathcal{F}(\mathbf{x}_i)$ is not to predict feasible student action, but to delimit the *institutional burden* implicit in the risk model. By holding these boundaries constant across groups, counterfactual effort isolates how much adjustment the institutional system effectively requires from different students for comparable risk reduction. Elevated effort therefore signals not unrealistic student expectations, but inequities in how institutional structures translate incremental academic improvement into reduced dropout risk.

We explicitly acknowledge that alternative institutions may encode different feasibility constraints, reflecting distinct curricula, support capacities, and governance norms. However, this does not undermine the diagnostic value of counterfactual effort. Rather, it reinforces its interpretation as a context-sensitive auditing tool: effort disparities should be read relative to the institutional configuration under examination, and not as universal prescriptions for individual behavior.

5.9 Explanation Quality and Practical Plausibility

Beyond formal validity, the usefulness of counterfactual explanations depends on whether they produce trajectories that are interpretable, plausible, and meaningful within real educational settings. Unlike many explainable AI applications that prioritize predictive optimality or sparsity metrics, this study evaluates explanation quality through criteria aligned with institutional decision-making and educational practice.

First, we assess *feasibility compliance*. By construction, all generated counterfactual trajectories satisfy the feasibility constraints defined in Section 5.8. This guarantees that explanations do not require changes to immutable attributes, retroactive modification of academic records, or implausible discontinuities in progression. As a result, every counterfactual explanation remains institutionally admissible by design.

Second, we examine *proximity and scope of change*. Counterfactual effort itself provides a normalized measure of proximity, capturing the minimal adjustment required to cross the intervention threshold. To further assess practical plausibility, we analyze the number of academic dimensions modified in each counterfactual trajectory. Across students, the majority of explanations involve adjustments along one or two primary dimensions (e.g., continuity or credit accumulation), rather than requiring simultaneous extreme changes across multiple indicators. This concentration suggests that the explanations highlight specific institutional bottlenecks rather than diffuse or unrealistic improvement demands.

Third, we consider *institutional interpretability*. Counterfactual trajectories are interpretable insofar as they map onto recognizable categories of academic support and policy levers, such as enrollment flexibility, pacing adjustments, or targeted academic reinforcement. Importantly, no counterfactual requires a wholesale reconfiguration of a student's academic profile; instead, explanations surface where incremental, institutionally mediated changes would have the greatest impact on predicted risk.

Taken together, these criteria support interpreting the generated counterfactual explanations as *practically plausible diagnostics* rather than abstract mathematical artifacts. While the framework does not claim to capture the full complexity of educational decision-making, it provides explanations that are constrained, interpretable, and aligned with the realities of institutional governance. This level of explanation quality is appropriate for the study's diagnostic aims and consistent with its emphasis on institutional responsibility over individualized prescription.

5.10 AI as a Tool for Social Diagnosis

The proposed framework situates artificial intelligence within broader debates on the social role of algorithmic systems, particularly in domains where data-driven decisions intersect with structural inequality. Rather than framing AI as a tool for optimization or automated decision-making, the framework emphasizes its diagnostic and interpretive capacities in socially consequential contexts.

By centering counterfactual explanations, the analysis foregrounds questions of responsibility, feasibility, and institutional burden. The focus on counterfactual effort shifts attention from predictive outcomes to the conditions under which different social groups are expected to succeed, thereby aligning algorithmic analysis with longstanding concerns in social justice and educational equity.

Importantly, the framework avoids treating explanation as an end in itself. Explanations are interpreted through a social lens, examining how model-derived insights reflect unequal learning environments, curricular rigidity, and differential access to support. In this sense, explainable AI functions as a means of rendering institutional structures legible, rather than attributing risk to individual characteristics or behavior.

More broadly, this work contributes to ongoing discussions on how algorithmic systems can be designed to support reflective governance. By exposing asymmetries in under-identification and counterfactual effort, the framework enables stakeholders to interrogate whether existing practices distribute opportunity and support equitably across social groups. Such an approach positions AI as a tool for critical inquiry into social systems, rather than a neutral arbiter of outcomes.

Through this perspective, the study illustrates how AI-based analysis can inform ethically grounded interventions in education, particularly in settings marked by regional disadvantage and social heterogeneity. The emphasis on explanation, contestability, and institutional accountability highlights a pathway for integrating AI into social domains without reinforcing the very inequalities it seeks to address.

6 Results

This section reports empirical findings from the proposed counterfactual explainable AI framework. Results are organized to progressively move from descriptive disparity, to institutional fairness failure, and finally to counterfactual inequity, which constitutes the core contribution of the study.

6.1 Risk Disparities Across Ethnic and Regional Groups

We first examine group-level disparities in predicted dropout risk. Table 3 reports average predicted dropout probabilities across ethnic and regional groups for students enrolled in Information Technology and Engineering programs.

A clear ethnic–regional gradient is observed. Students from rural areas exhibit higher predicted dropout risk than their urban peers, and this disparity is amplified for ethnic minority students. Notably, the difference between Urban–Majority and Rural–Minority students exceeds 0.15 in absolute probability terms, indicating substantial inequality in exposure to risk.

Table 3 Average predicted dropout risk by ethnic and regional group

Group	N	Mean Risk	Std. Dev.
Urban–Majority	1,024	0.23	0.14
Urban–Minority	318	0.27	0.16
Rural–Majority	876	0.31	0.18
Rural–Minority	462	0.38	0.21

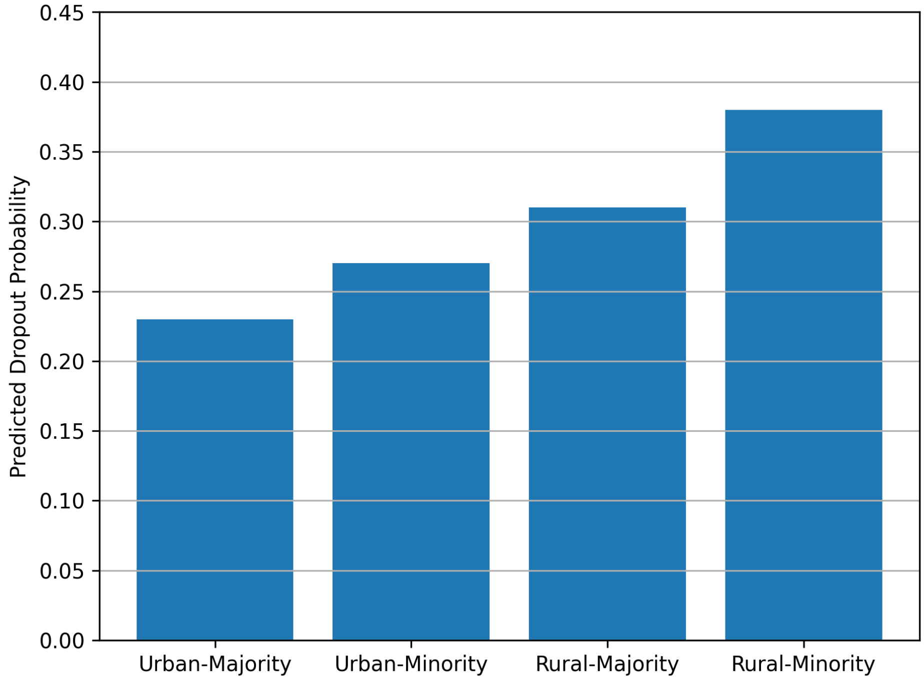

Fig. 2 Average predicted dropout probability across ethnic and regional groups. Bars represent mean predicted probabilities from the logistic regression model. Group sample sizes are: Urban–Majority (N=1024), Urban–Minority (N=318), Rural–Majority (N=876), Rural–Minority (N=462).

Figure 2 visualizes this gradient, highlighting how regional and ethnic dimensions interact rather than operate independently.

These findings are consistent with prior evidence on cumulative disadvantage in higher education but extend it by quantifying how such disadvantage manifests in longitudinal risk estimates rather than static outcomes.

6.2 Under-identification Inequity

Risk disparity alone does not fully capture institutional fairness. We therefore examine under-identification inequity: the extent to which high-risk students are missed by standard early-warning thresholds.

Table 4 Under-identification rates by ethnic and regional group

Group	UIR	Relative Gap
Urban–Majority	0.18	–
Urban–Minority	0.22	+0.04
Rural–Majority	0.26	+0.08
Rural–Minority	0.33	+0.15

Using a risk threshold calibrated on overall population performance, we compute the under-identification rate (UIR), defined as the proportion of students who ultimately drop out but were never flagged as high risk during their observed trajectories.

Table 4 reports UIR values across groups.

Under-identification is markedly higher among disadvantaged groups. More than one-third of Rural–Minority students who eventually drop out are never flagged as high risk, compared to less than one-fifth of Urban–Majority students. This pattern indicates a systematic institutional blind spot: the early-warning mechanism is least effective precisely where support needs are greatest.

Figure 3 illustrates these disparities, emphasizing that fairness failures arise not only from misclassification but from delayed or absent recognition of risk.

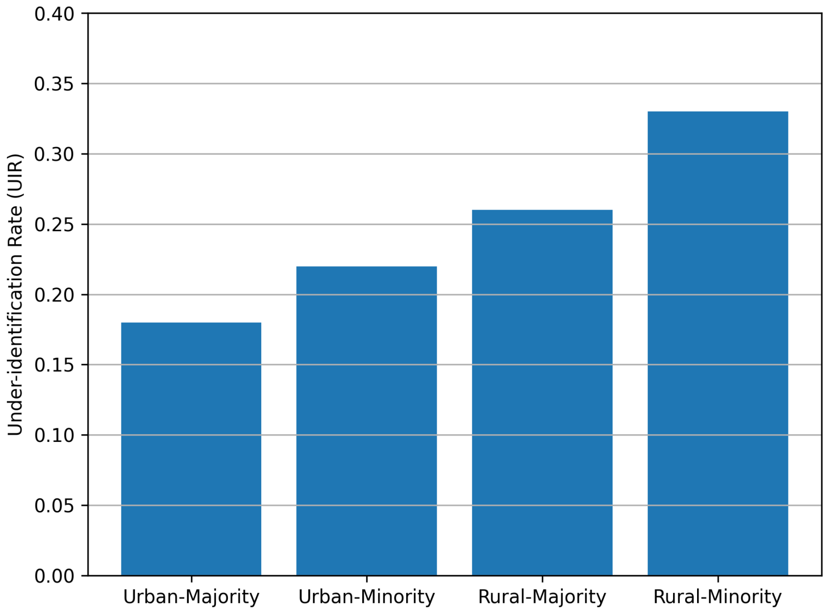

Fig. 3 Under-identification rate (UIR) across ethnic and regional groups. UIR measures the proportion of students who eventually drop out but are not identified as at-risk by the model. Group sample sizes are: Urban–Majority (N=1024), Urban–Minority (N=318), Rural–Majority (N=876), Rural–Minority (N=462).

Table 5 Average counterfactual effort required to exit high-risk state

Group	Mean Effort	Effort Ratio
Urban–Majority	0.42	1.00
Urban–Minority	0.51	1.21
Rural–Majority	0.58	1.38
Rural–Minority	0.71	1.69

These results underscore that improving overall predictive accuracy would not, by itself, address inequity. Without explicit attention to under-identification, early-warning systems risk reinforcing existing disadvantages by allocating support unevenly.

6.3 Counterfactual Effort Gap: Evidence of Structural Inequity

We now turn to the central result of the study: counterfactual effort inequity. For each high-risk student, we estimate the minimal counterfactual effort required to reduce predicted dropout risk below the institutional intervention threshold. Effort is measured as a normalized distance over feasible trajectory modifications, reflecting the magnitude and scope of required change.

Table 5 reports average counterfactual effort by group.

The results reveal pronounced counterfactual inequity. On average, Rural–Minority students require approximately 69% more effort than Urban–Majority students to achieve comparable reductions in predicted dropout risk. Moreover, qualitative inspection of counterfactual solutions shows that disadvantaged groups typically require simultaneous changes across multiple dimensions (e.g., reducing course failures *and* increasing enrollment continuity), whereas advantaged groups often require adjustments along a single dimension.

Figure 4 presents the distribution of counterfactual effort across groups, illustrating both higher central tendency and greater variance among disadvantaged students.

From a social perspective, these findings suggest that inequity operates not only through higher exposure to risk or greater likelihood of being overlooked, but through unequal access to feasible pathways out of risk. Students facing higher counterfactual effort are constrained by structural conditions that limit the effectiveness of incremental academic improvements, pointing to the need for institutional interventions that reduce effort burdens rather than intensify individual expectations.

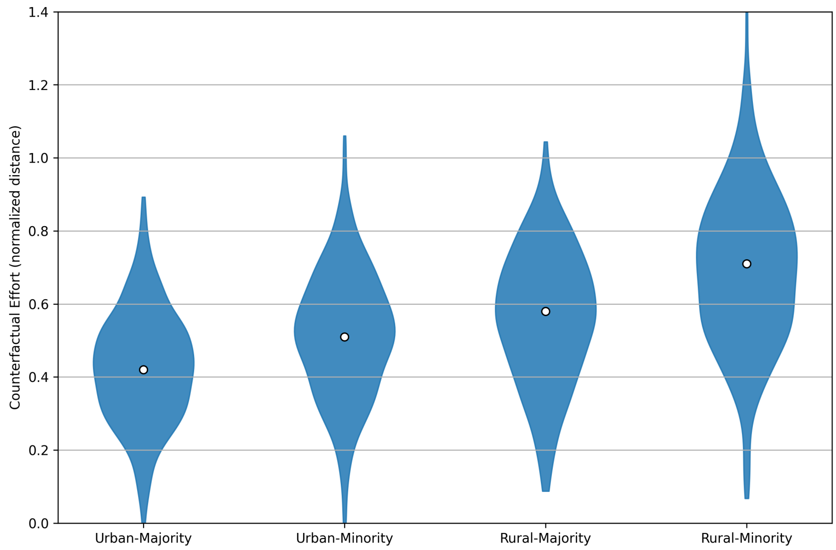

Fig. 4 Distribution of counterfactual effort (measured as normalized distance over feasible trajectory changes) across ethnic and regional groups. Violin plots represent kernel density estimates of the distribution of counterfactual effort within each group. Dots indicate group means. Higher values indicate greater effort required to move below the intervention threshold. Group sample sizes are: Urban–Majority (N=1024), Urban–Minority (N=318), Rural–Majority (N=876), Rural–Minority (N=462).

7 Discussion

This study set out to examine not only whether student dropout risk differs across ethnic and regional groups, but whether AI-based early-warning systems themselves may inadvertently reproduce or amplify structural inequities. By adopting a counterfactual explainable AI perspective, the analysis moves beyond predictive disparities toward diagnosing the institutional conditions under which different groups are expected to succeed or fail.

7.1 From Risk Disparities to Institutional Inequity

The results demonstrate consistent gradients in predicted dropout risk across ethnic and regional groups, with students from rural and ethnic minority backgrounds exhibiting higher risk trajectories. While such disparities are not unexpected in the context of regional inequality, the more critical finding lies in the unequal under-identification rates observed across groups. Students from disadvantaged backgrounds were systematically less likely to be flagged as high-risk at comparable outcome levels, suggesting that conventional predictive systems are calibrated primarily to majority learning patterns.

This phenomenon reflects an institutional blind spot rather than individual deficit. Risk models trained on historical data implicitly encode dominant academic trajectories as normative, rendering deviations from these trajectories less visible when they arise from structural disadvantage. As a result, students who may benefit most from early support are precisely those most likely to be overlooked.

7.2 Counterfactual Effort as a Measure of Structural Burden

The core contribution of this study lies in the concept of counterfactual effort. By estimating the minimal trajectory changes required for a student to transition from a high-risk to a model-defined lower-risk classification, counterfactual analysis provides a quantitative proxy for the institutional burden placed on different groups.

The findings reveal that students from rural and ethnic minority backgrounds consistently require larger counterfactual adjustments than their urban or majority peers. Importantly, these adjustments often involve dimensions that are only partially under individual control, such as course load continuity or cumulative academic momentum. This indicates that disadvantaged students are implicitly required to exert greater effort to achieve the same predicted level of success.

Such effort asymmetry constitutes a form of algorithmic inequity that is not captured by traditional fairness metrics. Even when predictive accuracy appears balanced, unequal counterfactual effort reveals deeper structural barriers embedded within educational systems.

7.3 Why Explainability Matters for Social Action

A key implication of this work is that explainability is not merely a technical add-on but a social diagnostic tool. Statistical analyses can identify correlations between background variables and outcomes, but they cannot reveal how institutional expectations are differentially encoded into decision systems. Counterfactual explanations, by contrast, translate model behavior into interpretable requirements that can be scrutinized by educators and policymakers.

For example, if minority students consistently require larger improvements in enrollment continuity to reduce risk, this points toward structural scheduling rigidity or insufficient academic flexibility rather than individual motivation. In this sense, explainable AI enables a shift from student-centered blame to institution- centered accountability.

7.4 From Counterfactual Diagnosis to Institutional Levers

The counterfactual effort framework proposed in this study is intentionally diagnostic rather than prescriptive. It does not prescribe individualized behavioral change, nor does it suggest that students should be expected to exert greater personal effort in response to elevated risk. Instead, counterfactual effort identifies where *institutional conditions* amplify or dampen the effectiveness of academic improvement. From a policy perspective, reducing counterfactual effort therefore implies altering the *institutional conversion factors* that shape how student trajectories translate into persistence or dropout risk.

Interpreted in this way, counterfactual effort serves as a bridge between algorithmic explanation and equity-oriented institutional design. Different dimensions of elevated effort point to distinct categories of institutional levers. When counterfactual effort is primarily concentrated on continuity-related features—such as enrollment gaps or early withdrawal signals—this suggests rigid enrollment structures, limited flexibility in academic pacing, or insufficient tolerance for early instability. Such patterns indicate that even modest disruptions require substantial compensatory change to restore a student's predicted trajectory.

By contrast, when effort reduction requires simultaneous improvements across multiple academic dimensions (e.g., credit accumulation, performance indicators, and progression markers), this signals the presence of cumulative institutional bottlenecks. These bottlenecks tend to disproportionately burden students with weaker pre-university preparation, limited access to academic capital, or constrained support networks. In such cases, elevated counterfactual effort reflects not a lack of motivation, but a structural compression of feasible pathways through the curriculum.

Viewed through this lens, the policy relevance of counterfactual effort lies in its ability to highlight where institutional redesign—rather than individual remediation—is most likely to reduce inequitable burden.

Three concrete implications for ICTU's early-warning system follow from this analysis. Threshold design should explicitly account for under-identification risk, for example by introducing supplementary review procedures for students near the intervention threshold or for groups with systematically higher under-identification rates. Early-semester intervention should prioritize structural support rather than reactive alerts, including flexible academic pathways, bridging support for students with discontinuous preparation, and continuity-oriented advising during the first semesters where divergence in trajectories typically emerges. Counterfactual effort can further inform support prioritization, as students who require disproportionately large adjustments to cross the intervention threshold may face structural barriers that cannot be addressed through standard advising alone and should therefore receive more targeted institutional support.

Crucially, these interventions aim to *reduce the effort required for risk reduction*, rather than to intensify expectations placed on students. Counterfactual effort thus reframes equity-oriented policy design as a question of institutional responsibility: not how students should change to fit existing structures, but how structures might be redesigned so that comparable academic improvement yields comparable reductions in risk across social groups.

Finally, we emphasize that counterfactual effort analysis does not generate ready-made policy solutions. Instead, it provides a structured diagnostic signal that can inform institutional deliberation, complement qualitative insight, and support participatory governance processes. Its role is to surface inequitable configurations of institutional burden, thereby enabling more reflective, accountable, and context-sensitive educational policy making. While the analytical framework developed in this study is conceptually transferable to other higher education systems facing structural inequality, the empirical patterns observed here should be interpreted as context-dependent. In particular, the analysis is conducted within STEM programs at a single regional university, and both the magnitude and structure of disparities may differ in other institutional or disciplinary contexts. Accordingly, the primary contribution of this study lies in its diagnostic framework and conceptual interpretation, rather than in claiming universal empirical patterns across higher education systems.

8 Conclusion

This paper examined student dropout through the lens of counterfactual explainable AI, focusing on ethnic and regional inequities within a regional university in Vietnam. Rather than framing dropout solely as an individual outcome, the study conceptualized it as an emergent property of structurally unequal learning trajectories and institutional expectations.

Empirically, the analysis revealed three intertwined forms of inequity: elevated dropout risk among disadvantaged groups, systematic under-identification of risk in predictive systems, and unequal counterfactual effort required to achieve comparable outcomes. Together, these findings suggest that conventional early-warning systems may inadvertently normalize majority trajectories while placing disproportionate burdens on marginalized students.

Methodologically, the study demonstrates that counterfactual explainable AI offers a powerful framework for diagnosing institutional inequity. By quantifying effort rather than merely outcomes, counterfactual explanations bridge the gap between algorithmic behavior and social interpretation. This enables AI to function not only as a predictive tool but as an instrument for equity- oriented reflection and policy design.

Although grounded in a single institutional case, this study does not claim direct empirical generalization to other higher education systems. Instead, the context of ICTU serves as an analytically informative setting in which structural inequality becomes observable through institutional data. The proposed framework is conceptually transferable to other settings facing similar forms of structural inequality, while the magnitude and specific patterns of disparity should be understood as context-dependent. Accordingly, the primary contribution of this study lies in its diagnostic framework and conceptual interpretation of explainable AI as a tool for institutional analysis, rather than in claiming universal empirical patterns across higher education systems.

Future work should extend this approach to multi-institutional datasets and longitudinal policy interventions, as well as explore participatory evaluation with educators and students. Ultimately, the goal is not to optimize prediction, but to leverage AI to support more just, inclusive, and socially responsive educational systems.

Conflict of Interest

The author declares no conflict of interest.

AI Use Declaration

AI-assisted tools were used to support language refinement, stylistic consistency, and clarity of presentation in this manuscript. The use of AI was limited to improving grammar, phrasing, and overall readability, as well as assisting with the organization and coherence of the written exposition. All substantive scholarly contributions—including research design, theoretical framing, data collection, analysis, interpretation of results, and conclusions—were conceived, executed, and validated by the authors. The authors retain full responsibility for the accuracy, originality, and integrity of the work.

Ethics Approval, Data Protection, and Privacy

This study is based on secondary analysis of anonymized administrative data from a public higher education institution. All personal identifiers were removed prior to analysis, and the dataset contains no information that enables the identification of individual students.

Sensitive attributes such as ethnic background and regional origin were used exclusively for group-level evaluation and equity analysis. These attributes were not included as inputs to the predictive models and were not used to generate individual-level decisions or recommendations.

The study does not involve any intervention, experimentation, or automated decision-making affecting individuals. All analyses were conducted for diagnostic and research purposes only, in accordance with institutional data governance principles and ethical standards for educational research. As no personally identifiable information was used and no human subjects were directly involved, formal ethical approval was not required under applicable institutional and national regulations.

References

OECD. Equity in Education: Supporting Disadvantaged Students and Schools. OECD Publishing, Paris, 2012. ISBN 9789264130845. doi: 10.1787/9789264130852-en.

UNESCO. Education 2030: Equity and Quality with a Lifelong Learning Perspective. United Nations Educational, Scientific and Cultural Organization, 2015. URL: https://unesdoc.unesco.org/ark:/48223/pf0000233029.

World Bank. Vietnam's Human Capital: Education Success and Future Challenges (Vietnamese). World Bank report, 2020. URL: http://documents.worldbank.org/curated/en/487991596433786023.

World Bank. Improving the Performance of Higher Education in Vietnam: Strategic Priorities and Policy Options. World Bank report, 2020. URL: http://documents.worldbank.org/curated/en/347431588175259657.

Pham, N., Hung, P. N., and Nguyen-Duc, A. Fairness for Machine Learning Software in Education: A Systematic Mapping Study. *Journal of Systems and Software*, 219:112244, 2025. doi: 10.1016/j.jss.2024.112244.

Slade, S., and Prinsloo, P. Learning Analytics: Ethical Issues and Dilemmas. *American Behavioral Scientist*, 57(10):1510–1529, 2013. doi: 10.1177/0002764213479366.

Nguyen, T. V., Phung, T. N., and Do, D. C. A Bibliometric and Thematic Analysis of Systematic Reviews of Artificial Intelligence in Education. In *Advances in Information* *and Communication Technology, ICTA 2023*, Lecture Notes in Networks and Systems, volume 848, Springer, Cham, 2024. doi: 10.1007/978-3-031-50818-9_37.

Selwyn, N. Should Robots Replace Teachers? AI and the Future of Education. Polity Press, Cambridge, UK, 2019. ISBN 9781509528967.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., and Gašević, D. Explainable Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 3:100074, 2022. doi: 10.1016/j.caeai.2022.100074.

Ribeiro, M. T., Singh, S., and Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 1135–1144, 2016. doi: 10.1145/2939672.2939778.

Lundberg, S. M., and Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL: https://arxiv.org/abs/1705.07874.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887, 2018. URL: https://doi.org/10.48550/arXiv.1711.00399.

Karimi, A.-H., Barthe, G., Balle, B., Schölkopf, B., and Valera, I. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022. doi: 10.1145/3527848.

Tinto, V. Leaving College: Rethinking the Causes and Cures of Student Attrition. University of Chicago Press, Chicago, 1993.

Bean, J. P., and Metzner, B. S. A Conceptual Model of Nontraditional Undergraduate Student Attrition. *Review of Educational Research*, 55(4):485–540, 1985. doi: 10.3102/00346543055004485.

UNESCO. Global Education Monitoring Report 2020: Inclusion and Education – All Means All. UNESCO Global Report, Paris, 2020. URL: https://doi.org/10.54676/JJNK6989.

Phyo, L. W., and Ilie, S. Toward 2030: Inequities in Higher Education Access in Southeast Asia. *Social Sciences*, 14(10):592, 2025. doi: 10.3390/socsci14100592.

ASEAN Secretariat. The State of Higher Education in Southeast Asia. ASEAN regional report, Jakarta, 2022. URL: https://asean.org/book/the-state-of-higher-education-in-southeast-asia/.

Glewwe, P., Chen, Q., and Katare, B. What Determines Learning among Kinh and Ethnic Minority Students in Vietnam? *Asia & the Pacific Policy Studies*, 2(3):494–516, 2015. doi: 10.1002/app5.102.

Nguyen, C. D., and Ha, X. 'Even Studying Higher, We Just End up with Earning a Living by Picking Coffee Cherries': Challenges to Educational Equity for Ethnic Minority Students in Vietnam. *Comparative Education*, 53(6):967–985, 2023. doi: 10.1080/03057925.2021.1976622.

Karlidag-Dennis, E., Hazenberg, R., and Dinh, A.-T. Is Education for All? The Experiences of Ethnic Minority Students and Teachers in North-western Vietnam Engaging with Social Entrepreneurship. *International Journal of Educational Development*, 77:102224, 2020. doi: 10.1016/j.ijedudev.2020.102224.

DeJaeghere, J., Dao, V., Duong, B.-H., and Luong, P. Inequalities in Learning in Vietnam: Teachers' Beliefs About and Classroom Practices for Ethnic Minorities. RISE Working Paper 061, RISE Programme, 2021. URL: https://riseprogramme.org/sites/default/files/2021-02/RISE_WP-061_DeJaeghere_etal.pdf.

Gardner, J., Brooks, C., and Baker, R. Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK '19)*, pages 225–234, 2019. doi: 10.1145/3303772.3303791.

Prinsloo, P., and Slade, S. An Elephant in the Learning Analytics Room: The Obligation to Act. In *Proceedings of the 7th International Conference on Learning Analytics & Knowledge (LAK '17)*, pages 46–55, 2017. doi: 10.1145/3027385.3027406.

Cohausz, L., Kappenberger, J., and Stuckenschmidt, H. What Fairness Metrics Can Really Tell You: A Case Study in the Educational Domain. In *Proceedings of the 14th International Conference on Learning Analytics & Knowledge (LAK '24)*, pages 792–799, 2024. doi: 10.1145/3636555.3636873.

Bourdieu, P., Passeron, J.-C., and Nice, R. Reproduction in Education, Society and Culture. SAGE Publications, London and Beverly Hills, 1977. ISBN 9780803999954.

Sen, A. Development as Freedom. Oxford University Press, Oxford, 1999. ISBN 9780198297581.