FAIR ENSEMBLE LEARNING FOR IMBALANCED DATA STREAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning from data streams requires online updates and stable predictive performance. However, class imbalance, sensitive-group imbalance, and instance-level difficulty imbalance are widespread in real-world data streams, where they reshape online training signals and may induce or amplify unfair predictions. Their distinct fairness effects and the limitations of existing fairness-aware methods under multi-source imbalance remain insufficiently understood. We analyze 204 controlled stationary synthetic streams using a unified profile-based prequential protocol. The results reveal distinct effects of the three imbalance sources on online fairness and highlight the challenges of maintaining fairness under their combined influence. Existing fairness interventions also exhibit source-dependent limitations and can incur substantial accuracy costs. Guided by these findings, we propose instance re**W**eighting for fair **E**nsemble **L**earning under imba**L**anced data streams (WELL). WELL jointly reweights the incoming instance according to label–sensitive cells and moderates within-cell difficulty, with theoretical conditions for fairness improvement and bounds on accuracy loss. Across five real-world streams and 19 synthetic settings, WELL outperforms state-of-the-art fairness-aware streaming methods in the overall accuracy–fairness trade-off.

1 INTRODUCTION

Real-world high-stakes systems, such as credit assessment and online risk prediction, often operate in stationary streaming environments, where instances arrive sequentially, and models are updated online (Domingos & Hulten, 2000; Bifet et al., 2010a). Since these decisions may affect individual opportunities, resource allocation, and long-term group-level outcomes, unfair predictions may disadvantage certain individuals or groups during online deployment (Hardt et al., 2016; Liu et al., 2018; Lee et al., 2026). Fairness in data stream learning is therefore a critical problem.

Real-world streams also commonly exhibit class imbalance (CI), sensitive-group imbalance (SI), and instance-level difficulty imbalance (IDI) (Wang et al., 2018; Aguiar et al., 2024). CI and SI skew the representation of labels and sensitive groups, respectively. IDI can be characterized by how instances are distributed across a widely adopted taxonomy comprising safe, borderline, rare, and outlier categories (Napierala & Stefanowski, 2016; Smith et al., 2014; Brzezinski et al., 2021), which encompasses common difficulty patterns, including class overlap, isolation, and label noise. Under finite and potentially skewed stream prefixes, these differences in representation and instance difficulty reshape the training evidence available to an online learner (He & Garcia, 2009; Krawczyk, 2016; Aguiar et al., 2024; Gao et al., 2026). Under such imbalanced streams, online updates may induce or amplify unfair model behavior (Subramanian et al., 2021; Tarzanagh et al., 2023).

Despite extensive research on imbalanced stream learning and growing work on fair streaming methods (Aguiar et al., 2024; Gillen et al., 2018; Zhang & Ntoutsi, 2019; Iosifidis & Ntoutsi, 2020; Lammers et al., 2025; Lee et al., 2023; Ghosh et al., 2026), three gaps remain. First, a unified analysis separating CI, SI, and IDI effects across learners and fairness measures is missing, particularly for within-cell difficulty and its different types. Second, the conditions under which existing interventions preserve fairness without substantial accuracy loss remain unclear across imbalance sources. Third, to our knowledge, existing methods lack explicit joint treatment of these sources. These sources interact during training: increasing scarce-cell influence can also amplify unreliable instances, while prediction disparity can exceed that represented by data statistics. Joint treatment must therefore

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

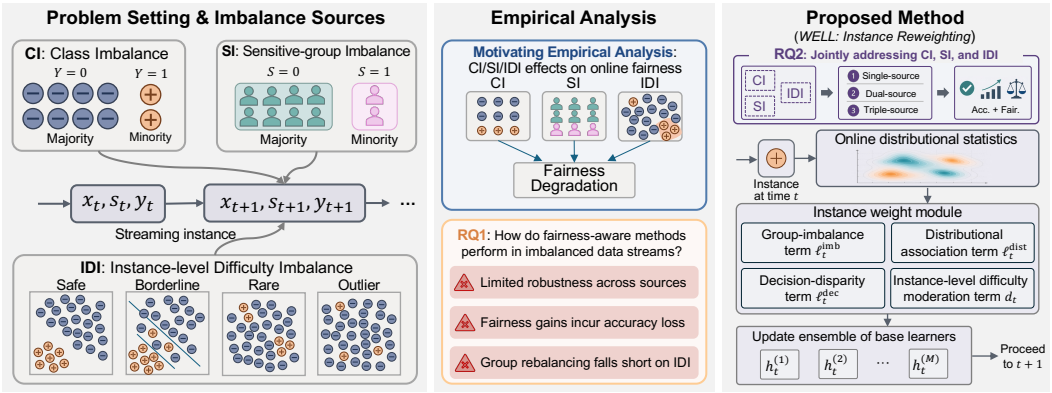


Figure 1: Overview of WELL under CI, SI, and IDI in data streams. WELL links empirical fairness degradation to online instance reweighting for robust accuracy–fairness trade-offs.

account for both the allocation and the reliability of training evidence. We first conduct a motivating empirical analysis of how CI, SI, and IDI affect online fairness. Building on this analysis, we address two research questions (RQs). **RQ1:** How do fairness-aware methods perform in imbalanced data streams? **RQ2:** How can online learning jointly address CI, SI, and IDI to improve the accuracy–fairness trade-off?

For this study, we establish a unified analysis framework with controlled stationary streams that vary CI, SI, and IDI independently or jointly. Since endpoint-only fairness evaluation can obscure disparities during online updates, we use a profile-based prequential protocol to track fairness over valid stream prefixes. We study 15 representative state-of-the-art online learners and six state-of-the-art fairness-aware streaming methods, then use the findings to develop instance re**W**eighting for fair **E**nsemble **L**earning under imba**L**anced data streams (WELL).

Our contributions are threefold. (i) We establish distinct empirical fairness patterns: CI reshapes error disparities, SI affects a broader set of fairness measures, and IDI effects depend on difficulty type. We also identify amplification of label-side disparity by online learners. (ii) We show that existing fairness interventions do not retain their benefits uniformly across imbalance sources, and distinguish accuracy costs from failure to handle within-cell difficulty. (iii) Guided by this evidence, WELL jointly adjusts training influence across label–sensitive cells and within each cell. Ablations support this combination, while theory links the weighting statistics to conditional SPD/DI improvement, bounds accuracy loss, and establishes within-cell difficulty attenuation.

2 Related Work

Learning from Imbalanced Data Streams. Learning from imbalanced data streams has been widely studied in online learning, with most work focusing on CI, where minority-class samples are sparsely observed and can be dominated by majority-class updates (He & Garcia, 2009; Krawczyk, 2016; Wang et al., 2018; Aguiar et al., 2024). Existing methods include data-level resampling, cost-sensitive updates, and ensemble-based learning (Wang et al., 2018; Aguiar et al., 2024; Gao et al., 2026). Instance-level difficulty has also been studied in imbalanced data streams, showing that borderline, rare, and outlier samples can induce highly variable learning signals (Napierala & Stefanowski, 2016; Smith et al., 2014; Brzezinski et al., 2021). However, this literature mainly treats imbalance as a predictive learning problem, with evaluation centered on accuracy performance. Its implications for fairness under sequential model updates remain much less understood.

Fair Streaming Learning. Fair streaming methods intervene through resampling (Ghosh et al., 2026; Lammers et al., 2025), model updates (Zhang & Ntoutsi, 2019; Lee et al., 2023), multi-objective learning (Pham et al., 2024), and ensembles (Iosifidis & Ntoutsi, 2020). FABBOO combines imbalance handling with prediction feedback, while CFSMOTE uses situation testing and oversampling. Factorized-reference reweighting also has established precedents (Kamiran & Calders, 2012). We investigate how the three imbalance sources affect fairness and these interventions within a uni-

fied controlled study. This evidence motivates WELL's joint treatment of cell-level correction and within-cell difficulty, whose complementary effects are evaluated separately.

3 PRELIMINARIES

Notation. We consider a stationary stream $\mathcal{D} = \{(x_t, s_t, y_t)\}_{t=1}^{T}$, a widely encountered yet challenging online setting (Domingos & Hulten, 2000; Bifet et al., 2010a; Gao et al., 2026), where $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$, $s_t \in \{0, 1\}$, and $y_t \in \{0, 1\}$ are the instance, sensitive attribute, and label. Samples follow a fixed unknown $P(X, S, Y)$. The learner predicts $\hat{y}_t = f_t(x_t, s_t)$ before observing y_t and then updates f_t; we write $g = (y, s) \in \mathcal{G} = \mathcal{Y} \times \mathcal{S}$ for a label–sensitive cell.

Definition 1 (Class Imbalance, CI (Fernández et al., 2018))**.** *CI refers to skewness in the label distribution $P(Y)$. For binary classification, $\rho_{\mathrm{CI}} = \min_{y \in \mathcal{Y}} P(Y = y)$. A smaller ρ_{CI} indicates stronger class imbalance.*

Definition 2 (Sensitive-group Imbalance, SI (Kamiran & Calders, 2012))**.** *SI refers to skewness in the sensitive attribute distribution $P(S)$. For binary sensitive groups, $\rho_{\mathrm{SI}} = \min_{s \in \mathcal{S}} P(S = s)$. A smaller ρ_{SI} indicates stronger sensitive-group imbalance.*

Definition 3 (Instance-level Difficulty Imbalance, IDI (Napierala & Stefanowski, 2016))**.** *IDI is skewness in local same-class support such that difficult, low-support instances occupy a disproportionate share of a stream. This widely observed heterogeneity is represented by the established safe/borderline/rare/outlier taxonomy in the real world (Napierala & Stefanowski, 2016; Smith et al., 2014; Brzezinski et al., 2021): safe instances are dense and separable, borderline instances reflect overlap or neighborhood disagreement, rare instances reflect sparse subconcepts, and outliers reflect isolation or strong label conflict, including possible label noise.*

4 EVALUATION PROTOCOL

Profile-based Fairness Evaluation. Window-level fairness scores collapse online behavior into a single endpoint value and may miss transient unfairness during model updates. Let ϕ be a base fairness functional and let $\mathcal{F} = \{f_t\}_{t \geq 1}$ denote the online predictor sequence. For a window ending at time t, define the prequential record prefix

$$\mathcal{R}_\tau^{\mathcal{F}}(t, w) = \left\{(s_u, y_u, \hat{y}_u^{\mathcal{F}})\right\}_{u=t-w+1}^{\tau}, \qquad \hat{y}_u^{\mathcal{F}} = f_u(x_u, s_u), \quad \tau \in \mathcal{V}_{t,w},$$

where $\mathcal{V}_{t,w} = \{t - w + 1, \ldots, t\}$ and each prediction is made before observing y_u. We define the fairness profile as the ordered vector

$$\mathbf{P}_\phi^{\mathcal{F}}(t, w) = \Big(\phi(\mathcal{R}_\tau^{\mathcal{F}}(t, w))\Big)_{\tau \in \mathcal{V}_{t,w}} \in \mathbb{R}^w.$$

For scalar comparison, we aggregate the profile by $\overline{\Phi}_\phi^{\mathcal{F}}(t, w) = \langle \boldsymbol{\lambda}_{t,w}, \mathbf{P}_\phi^{\mathcal{F}}(t, w) \rangle$, where $\boldsymbol{\lambda}_{t,w} \in \Delta^{w-1}$ is an exponentially decayed weight vector in our experiments and details are given in Appendix B.1. This scalar summary is used only for compact method comparison, while the profile vector itself preserves temporal information that is discarded by endpoint-only window evaluation. Let $E_\phi^{\mathcal{F}}(t, w) := \phi(\mathcal{R}_t^{\mathcal{F}}(t, w))$ denote the endpoint fairness score over the same window.

Proposition 1 (Endpoint evaluation as a profile projection)**.** *For any window ending at time t, $E_\phi^{\mathcal{F}}(t, w) = \pi_{\mathrm{last}}\Big(\mathbf{P}_\phi^{\mathcal{F}}(t, w)\Big)$, where π_{last} extracts the last coordinate of the profile. Moreover, for standard rate-based fairness metrics such as demographic parity difference, two valid prequential window traces can have the same endpoint score but different fairness profiles.*

Proposition 1 shows that endpoint-only evaluation is a degenerate projection of the profile and can therefore miss transient fairness changes within the window. Thus, the profile evaluates the temporal fairness behavior of the evolving online learner, whereas $\overline{\Phi}_\phi^{\mathcal{F}}(t, w)$ provides a compact scalar summary for method comparison. Proofs and valid-prefix details for ratio and conditional-rate metrics are given in Appendix B.1.

Evaluation Metrics. We report intersectional balanced accuracy (BA), averaging accuracy over observed (Y, S) cells, and Kappa for accuracy. SPD and DI are the main fairness measures, with all 22 measures defined in Appendix C. For binary S and $\hat{Y}$, $q_s = \mathbb{P}(\hat{Y} = 1 \mid S = s)$, $\mathrm{SPD} = |q_0 - q_1|$, $\mathrm{DI} = \left|1 - \min\left\{\frac{q_0}{q_1}, \frac{q_1}{q_0}\right\}\right|$. Smaller SPD and DI indicate better fairness. We analogously measure label-side disparity by $\Delta^{\mathrm{SPD}}_{\mathrm{data}} = |\mathbb{P}(Y = 1 \mid S = 0) - \mathbb{P}(Y = 1 \mid S = 1)|$, and compute all fairness results with the profile protocol above.

Datasets and imbalance scenarios. We evaluate 204 synthetic and five real-world streams. HYPERPLANE (Hulten et al., 2001) and RANDOMRBF (Bifet et al., 2010a) vary CI and SI under data-side fair and unfair conditions, while DIFFICULTY (Brzezinski et al., 2021) varies borderline, rare, and outlier composition. Appendix D specifies the grids and real-stream sources. Profile-matched synthetic–real comparisons support the findings' applicability beyond controlled generators (Appendix D.8).

5 MOTIVATING EMPIRICAL ANALYSIS

Experiment Setup. We evaluate 15 representative state-of-the-art online learning algorithms on controlled stationary streams. They include standard and imbalance-aware stream learners, but none optimizes fairness, enabling us to isolate how CI, SI, and IDI affect fairness through online dynamics. We construct stationary streams $\{\mathcal{D}^{(r)}\}_{r=1}^{R}$ with fixed distributions and controlled imbalance profiles. CI and SI are quantified by ρ_{CI} and ρ_{SI}, where smaller values indicate stronger imbalance, whereas IDI is controlled through the proportions of safe, borderline, rare, and outlier instances. We vary one imbalance source at a time and consider both fair and unfair data-side settings measured by $\Delta^{\mathrm{SPD}}_{\mathrm{data}}$. Full experimental details and additional findings are provided in Appendices E and F, respectively.

Finding 1: CI exposes algorithmic unfairness through group-wise error disparities. Figure 2a shows that stronger CI redistributes classification errors across sensitive groups. Degradation is clearest in FPR-, FDR-, APD-, and ER-related measures, whereas AOD and AAOD change less. SPD and DI respond differently because DI is ratio-based and more sensitive to small positive-rate changes. Difference- and ratio-based metrics should be interpreted jointly under CI.

Finding 2: SI broadly degrades group-based fairness in algorithmic decisions. Following the across-stream setting of Finding 1, Figure 2b compares stationary streams with fixed $\Delta^{\mathrm{SPD}}_{\mathrm{data}} = 0.25$ and varying SI degrees $\rho^{(r)}_{\mathrm{SI}}$. Stronger SI degrades group-based fairness more broadly than CI. SPD, DI, and FNR-, FPR-, FOR-, FDR-, APD-, and ER-related measures degrade for most learners, whereas individual and inequality-based measures respond less consistently. Thus, SI alters both group-wise prediction rates and error profiles.

Finding 3: Imbalance amplifies algorithmic unfairness inherited from biased data streams. The average amplification $\Delta^{\mathrm{SPD}}_{\mathrm{amp}} = \Delta^{\mathrm{SPD}}_{\mathrm{model}} - \Delta^{\mathrm{SPD}}_{\mathrm{data}}$ is often positive across 15 learners, particularly under stronger SI and rare- or outlier-dominated IDI. Amplification often grows with label-side disparity, indicating that imbalance can reinforce existing unfairness during sequential updates. Hence label-side statistics alone do not characterize current model disparity.

Finding 4: IDI severity amplifies fairness degradation, especially under rare and outlier dominance. Figure 3 shows that SPD generally increases with IDI severity, indicating worse fairness. $\rho_{\mathrm{IDI}} \in \{0, 0.2, 0.4, 0.6, 0.8\}$ denotes the proportion of the dominant difficulty type. Borderline-dominated streams change modestly, whereas rare- and outlier-dominated streams show sharp increases, particularly when $\rho_{\mathrm{IDI}} \geq 0.6$. These results indicate that sparse or unstable evidence from rare and outlier instances makes online updates susceptible to fairness degradation. Thus, IDI's fairness effect depends on both its severity and the difficulty type dominating the stream.

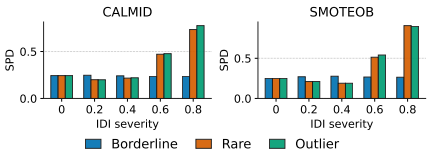
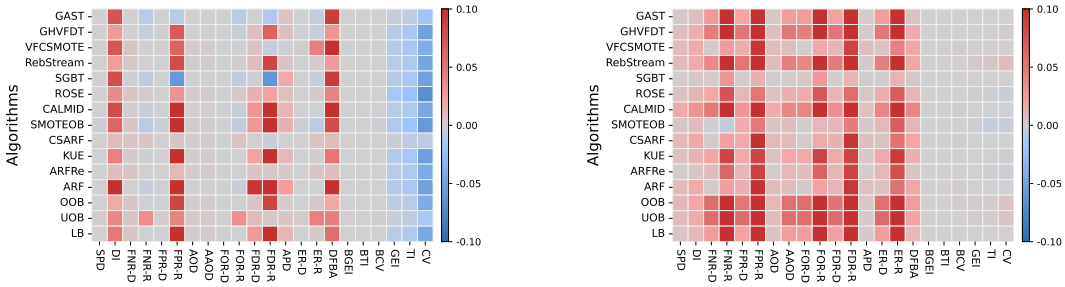


Figure 3: SPD under borderline-, rare-, and outlier-dominated IDI. Panels show representative algorithms; larger values indicate worse fairness. Appendix F.6 provides details.



(a) Only CI, $\Delta^{\mathrm{SPD}}_{\mathrm{data}} = 0.25$

(b) Only SI, $\Delta^{\mathrm{SPD}}_{\mathrm{data}} = 0.25$

Figure 2: Metric-wise fairness slopes under CI and SI. Red and blue indicate degradation and mitigation. Appendix E.3 gives the slope computation and more results.

6 RQ1: How Do Fairness-aware Methods Perform in Imbalanced Data Streams?

Experiment Setup. To answer RQ1, we evaluate six fairness-aware data-stream methods on controlled stationary streams, including methods with CI- or SI-aware mechanisms. Notably, existing fairness-aware data-stream methods have not explicitly considered IDI. We use the same stream construction, imbalance controls, and prequential protocol as the motivating analysis, varying CI, SI, and IDI separately while keeping other factors fixed. Guided by its findings, we report BA/Kappa for accuracy and SPD/DI for fairness, since difference- and ratio-based parity metrics can behave differently yet provide complementary evidence under imbalance. Full experimental results are provided in Appendix G.

Finding 5: Fairness interventions without imbalance modeling show limited robustness under imbalanced streams. Figure 4 shows that a method can achieve better fairness under one form of imbalance and substantially worse fairness under another. Lower SPD and DI values indicate better fairness. CFDS, for example, keeps SPD near 0.2–0.3 throughout the CI and SI settings, but its SPD exceeds 0.6 at $\rho_{\mathrm{IDI}} = 0.6$ and approaches 0.9 at $\rho_{\mathrm{IDI}} = 0.8$ under outlier-dominated IDI. This increase shows that CFDS becomes substantially less fair as outlier-dominated IDI strengthens. The relative fairness ordering of the methods also changes across imbalance forms and severity levels.

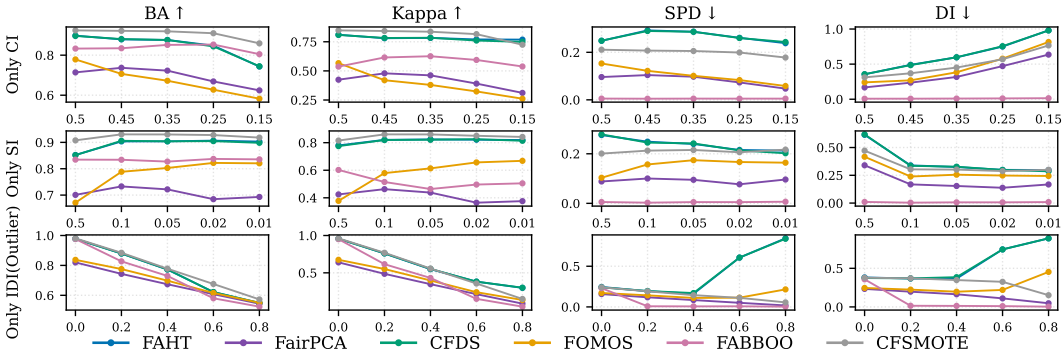


Figure 4: Accuracy and fairness of existing methods as CI, SI, or outlier-dominated IDI varies.

Finding 6: Existing imbalance-aware methods often improve algorithmic fairness at a substantial accuracy cost. Figure 4 shows that better fairness can require substantial accuracy loss. At $\rho_{\mathrm{SI}} = 0.01$, FABBOO is closest to the ideal SPD and DI value of zero, yet reaches only about 0.83 BA and 0.51 Kappa, whereas CFDS and CFSMOTE reach BA near 0.90 and Kappa above 0.80. Similar patterns under severe imbalance make joint accuracy–fairness evaluation essential.

Finding 7: IDI is not adequately addressed by group-level rebalancing in the evaluated settings. Figure 4 shows that group-level rebalancing cannot distinguish locally unreliable instances within a (Y, S) cell. With CI and SI fixed, increasing outlier prevalence from 0 to 0.8 reduces CFSMOTE's BA from about 0.98 to 0.57 and Kappa from 0.97 to 0.15. Together with the motivating analysis, this severe loss motivates instance-level reliability moderation alongside group correction.

7 WELL: Instance Reweighting for Fair Ensemble Learning

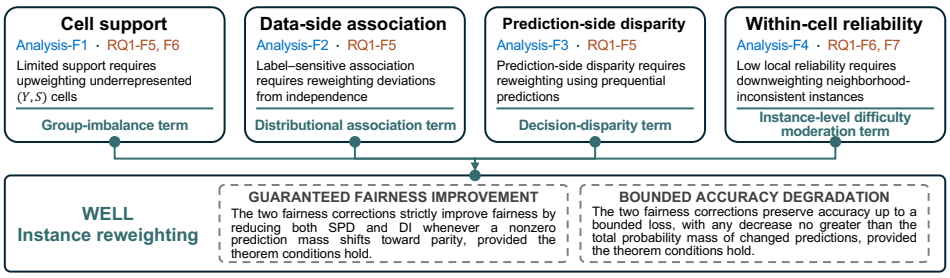


Figure 5: From empirical findings to WELL's design and conditional theoretical guarantees.

From empirical findings to WELL. The motivating analysis and RQ1 suggest three IDI-related challenges: (i) within-cell heterogeneity, (ii) degraded evidence in CI/SI-sparse cells, and (iii) uneven group-error responses that worsen accuracy–fairness trade-offs (Appendix I.1). Figure 5 links these findings to corrections for cell scarcity, label–sensitive association, prediction-side disparity, and within-cell difficulty. WELL's cell-level terms adjust influence across cells, while difficulty moderation distinguishes observations within each cell. Their combination supports scarce-cell correction without treating all its observations as equally reliable. Each arrival's weight uses current statistics and recent-neighborhood agreement (Algorithm 1). Theorem 1 establishes conditional fairness and accuracy properties.

Ensemble of base learners. At time t, WELL aggregates its M base learners, $f_t(x_t, s_t) = \mathrm{Ens}\big(h_t^{(1)}(x_t, s_t), \ldots, h_t^{(M)}(x_t, s_t)\big)$, and predicts $\hat{y}_t = f_t(x_t, s_t)$ before observing y_t. WELL then computes a canonical weight ω_t for the current instance. To preserve ensemble diversity, learner m receives $\omega_t^{(m)} = \mathcal{D}_m(\omega_t; \xi_t^{(m)})$, where $\mathcal{D}_m$ is a diversification operator and $\xi_t^{(m)}$ denotes learner-specific randomness or configuration (Oza & Russell, 2001; Wang et al., 2016). The update is

Algorithm 1: WELL

Require: Stream $\mathcal{D}$, ensemble $\mathcal{H}_1$, buffer $\mathcal{B}_1$
Require: $\beta, \gamma_{\mathrm{dist}}, \gamma_{\mathrm{dec}}$ (all 1 by default)
for $t = 1, \ldots, T$ **do**
  Predict $\hat{y}_t = f_t(x_t, s_t)$ and reveal y_t
  Estimate stream and prediction statistics from $\mathcal{I}_t$
  Compute $\ell_t^{\mathrm{imb}}, \ell_t^{\mathrm{dist}}, \ell_t^{\mathrm{dec}}$ and a_t^{dec}, d_t
  Set $\eta_t = \beta \ell_t^{\mathrm{imb}} + \gamma_{\mathrm{dist}} \ell_t^{\mathrm{dist}} + \gamma_{\mathrm{dec}} a_t^{\mathrm{dec}} \ell_t^{\mathrm{dec}}$
  Set $\omega_t = \exp(\eta_t) \cdot \chi(d_t)$
  Update each learner via $\mathcal{D}_m(\omega_t; \xi_t^{(m)})$ by Eq. equation 1
  Add $(x_t, s_t, y_t, \hat{y}_t)$ to statistics and $\mathcal{B}_t$
end for

$$h_{t+1}^{(m)} = \mathrm{Update}\Big(h_t^{(m)}; x_t, s_t, y_t, \omega_t^{(m)}\Big), \qquad m = 1, \ldots, M. \tag{1}$$

Here ω_t determines the instance's training influence, while $\mathcal{D}_m$ distributes it across the ensemble through the lightweight weight-perturbation scheme detailed in Appendix H.

Online distributional statistics. WELL estimates the distributional statistics from the observed stream. Before updating on instance t, it estimates label–sensitive cell frequencies, data-side laws over (Y, S), and decision-side laws over $(\hat{Y}, S)$, denoted by $\hat{p}_t(y, s)$, $\hat{p}_t(y)$, $\hat{p}_t(s)$, $\hat{p}_t^{\mathrm{pred}}(\hat{y}, s)$, $\hat{p}_t^{\mathrm{pred}}(\hat{Y} = \hat{y})$, and $\hat{\mathbb{P}}_t(\hat{Y} = 1 \mid S = s)$. The weight uses the revealed label to identify the current cell and assess neighbor agreement, while prediction remains test-before-train. Details are given in Appendix H.

Group-imbalance and parity terms. For the current label–sensitive cell (y_t, s_t), WELL uses three log-weight terms:

$$\ell_t^{\mathrm{imb}} = \log \frac{\max_{(y,s)\in\mathcal{G}} \hat{p}_t(y,s)}{\hat{p}_t(y_t,s_t)}, \tag{2}$$

$$\ell_t^{\mathrm{dist}} = \log \frac{\hat{p}_t(y_t)\hat{p}_t(s_t)}{\hat{p}_t(y_t,s_t)} - \frac{1}{|\mathcal{G}|}\sum_{(y,s)\in\mathcal{G}} \log \frac{\hat{p}_t(y)\hat{p}_t(s)}{\hat{p}_t(y,s)}, \tag{3}$$

$$\ell_t^{\mathrm{dec}} = \log \frac{\hat{p}_t^{\mathrm{pred}}(\hat{Y}=y_t)\hat{p}_t(s_t)}{\hat{p}_t^{\mathrm{pred}}(\hat{Y}=y_t,S=s_t)} - \frac{1}{|\mathcal{G}|}\sum_{(y,s)\in\mathcal{G}} \log \frac{\hat{p}_t^{\mathrm{pred}}(\hat{Y}=y)\hat{p}_t(s)}{\hat{p}_t^{\mathrm{pred}}(\hat{Y}=y,S=s)}. \tag{4}$$

Equation equation 2 upweights scarce (Y, S) cells. Equation equation 3 centers the log deviation of the observed joint mass from the factorized reference $\hat{p}_t(Y)\hat{p}_t(S)$, thereby isolating label–sensitive association from marginal scarcity. Equation equation 4 applies the same principle to historical prequential predictions, with $\hat{Y} = y_t$ as the supervised target outcome. Let $\widehat{\Delta}^{\mathrm{DP}}_{\mathrm{dec},t}$ denote the historical difference between the groups' positive prediction rates. A fixed bounded, nondecreasing mapping ψ with $\psi(0) = 0$ gives $a_t^{\mathrm{dec}} = \psi(\widehat{\Delta}^{\mathrm{DP}}_{\mathrm{dec},t})$, so the decision-side contribution varies continuously with prediction disparity and vanishes when the rates coincide. Appendix H gives the streaming estimates.

Instance-level difficulty moderation. To avoid amplifying unreliable rare or outlier-like instances, WELL estimates local label consistency from the recent buffer $\mathcal{B}_t$. For neighborhood size k,

$$\mathcal{K}_t = \mathrm{kNN}_k(x_t;\mathcal{B}_t), \qquad r_t = \frac{1}{k}\sum_{i\in\mathcal{K}_t} \mathbf{1}(y_i = y_t), \qquad d_t = 1 - r_t. \tag{5}$$

Eq. equation 5 defines r_t as local agreement with the label and d_t as a finite-buffer difficulty proxy. A larger d_t indicates lower local consistency, so inconsistent instances receive reduced influence rather than amplification by rarity- or parity-based terms. The recent-buffer kNN provides a simple, effective non-parametric learning mechanism: it learns the local feature–label relationship from observed instances and converts it into a reliability estimate that is updated as the buffer evolves.

Instance weight. WELL combines the above components through an exponential log-weight:

$$\omega_t = \exp\big(\beta\ell_t^{\mathrm{imb}} + \gamma_{\mathrm{dist}}\ell_t^{\mathrm{dist}} + \gamma_{\mathrm{dec}}a_t^{\mathrm{dec}}\ell_t^{\mathrm{dec}}\big)\cdot\chi(d_t), \tag{6}$$

Here β, γ_{dist}, and γ_{dec} scale group imbalance, distributional association, and decision disparity, respectively, while $\chi(d_t)$ moderates difficulty. Comparisons with arithmetic- and geometric-mean alternatives support this multiplicative combination empirically (Appendix I.1), complementing the conditional analysis of its log-additive structure below.

Theoretical justification. The exponential form makes the roles of the three outer coefficients explicit. Fix the pre-update history and compare WELL with a matched rule that retains the same group-imbalance and difficulty terms but removes the two fairness terms. For cell $g = (y, s)$, define

$$\begin{aligned} \Gamma_t(g) &= \gamma_{\mathrm{dist}}\ell_t^{\mathrm{dist}}(g) + \gamma_{\mathrm{dec}}a_t^{\mathrm{dec}}\ell_t^{\mathrm{dec}}(g), \\ L_t^{\mathrm{full}}(g) &= \beta\ell_t^{\mathrm{imb}}(g) + \Gamma_t(g), \qquad L_t^{\mathrm{matched}}(g) = \beta\ell_t^{\mathrm{imb}}(g). \end{aligned} \tag{7}$$

Thus $\omega_t^{\mathrm{full}} = \chi(d_t)e^{L_t^{\mathrm{full}}}$ and $\omega_t^{\mathrm{matched}} = \chi(d_t)e^{L_t^{\mathrm{matched}}}$. With β fixed, its effect remains in both rules and cancels only when dividing their relative cell ratios. Appendix B.2 gives the full proof.

Lemma 1 (Cell-ratio and threshold response)**.** *Fix $\mathcal{I}_t$, write $P_t(\cdot) = P(\cdot \mid \mathcal{I}_t)$, and let $Z_t^\chi = \mathbb{E}_{P_t}[\chi(d_t)] > 0$. Define the common difficulty law $Q_t^\chi(A) = \mathbb{E}_{P_t}[\mathbf{1}\{(X,Y,S)\in A\}\chi(d_t)]/Z_t^\chi$ and posterior $\eta_{t,s}^\chi(x) = Q_t^\chi(Y=1 \mid X=x, S=s)$. For $r \in \{\mathrm{matched}, \mathrm{full}\}$, let $\mathcal{W}_t^r$ have density proportional to $\chi(d_t)e^{L_t^r(Y,S)}$ relative to P_t. For relevant cells g, g', normalization gives*

$$\log \frac{\mathcal{W}_t^{\mathrm{full}}(g)/\mathcal{W}_t^{\mathrm{full}}(g')}{\mathcal{W}_t^{\mathrm{matched}}(g)/\mathcal{W}_t^{\mathrm{matched}}(g')} = \Gamma_t(g) - \Gamma_t(g').$$

Set $w_{ys}^r = e^{L_t^r(y,s)}$. Pointwise minimization of weighted binary log loss under Q_t^χ predicts positive exactly when $\eta_{t,s}^\chi(x) \geq \tau_{t,s}^r := w_{0s}^r/(w_{0s}^r + w_{1s}^r)$. Since $\log[(w_{1s}^{\mathrm{full}}/w_{0s}^{\mathrm{full}})/(w_{1s}^{\mathrm{matched}}/w_{0s}^{\mathrm{matched}})] = \Gamma_t(1,s) - \Gamma_t(0,s)$, the fairness corrections induce group-specific threshold shifts while preserving feature-dependent difficulty moderation.

Theorem 1 (Conditional parity change and accuracy bound)**.** *Let H and L be the higher- and lower-positive-rate groups of the matched-rule population response, with rates $r_H \geq r_L > 0$. Assume positive cells and weights, realization of weighted population minimizers, parity-correcting threshold shifts, and no parity crossing. If masses $m_H, m_L \geq 0$ switch from $1 \to 0$ in H and $0 \to 1$ in L, respectively, with $r_H - m_H \geq r_L + m_L > 0$, then, writing B and F for the matched and full rules,*

$$\mathrm{SPD}_B - \mathrm{SPD}_F = m_H + m_L, \qquad \mathrm{DI}_B - \mathrm{DI}_F = \frac{r_H m_L + r_L m_H}{r_H(r_H - m_H)}, \\ \mathrm{Acc}_F \geq \mathrm{Acc}_B - P(S = H)m_H - P(S = L)m_L. \tag{8}$$

Both SPD and DI decrease strictly when $m_H + m_L > 0$.

Theorem 1 quantifies conditional SPD/DI improvement and bounds accuracy loss. Appendix B.2 links the response to WELL's statistics, refines the accuracy bound using label posteriors, and proves within-cell difficulty attenuation, explaining the complementary roles of cell-level fairness correction and instance-level moderation.

8 RQ2: How Can Online Learning Jointly Address CI, SI, and IDI to Improve the Accuracy–Fairness Trade-Off?

Experiment Setup. To answer RQ2, we compare WELL with the same fairness-aware methods used in RQ1. The evaluation covers 19 representative settings, spanning single-source, pairwise, and joint CI–SI–IDI imbalance, together with five real-world data streams. BA and Kappa measure accuracy, while SPD and DI measure fairness. We adopt the widely used geometric-mean aggregation (Zhang et al., 2023) to compute GM_4 for joint accuracy–fairness evaluation (Zhang et al., 2025), with lower values indicating better trade-offs. Computation details are provided in Appendix E.7. We also evaluate component ablations, coefficient sensitivity, difficulty-proxy alignment, and fusion rules. Full results are in Appendix I.

WELL maintains a strong accuracy–fairness trade-off across diverse CI/SI/IDI combinations. WELL ranks first in fairness and GM_4 and fourth in accuracy (Figure 6). On Adult, it improves BA, Kappa, SPD, and DI over FABBOO, demonstrating simultaneous accuracy and fairness gains (Appendix I.3). Joint CI+SI+IDI results also show accuracy gains at low disparity, although benefits vary across settings (Appendix I.2). The metric-wise results complement the rank comparison.

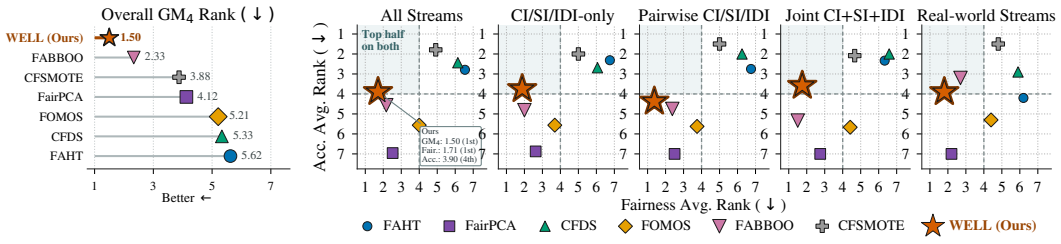


Figure 6: Accuracy–fairness rank trade-off across imbalance settings. Each point reports average accuracy rank and fairness rank, where lower is better on both axes, and upper-left region is preferred.

WELL components play complementary roles. In Table 1, removing both fairness terms, $\ell_t^{\mathrm{fair}} = \{\ell_t^{\mathrm{dist}}, \ell_t^{\mathrm{dec}}\}$, yields the best accuracy but worst fairness and GM_4 ranks. Removing ℓ_t^{dec} causes the largest fairness loss among single-term ablations.

Difficulty moderation improves accuracy and fairness on average under IDI. Figure 7 shows aggregate accuracy and fairness gains under IDI, with mixed effects otherwise. Difficulty moderation improves all four metrics on rare- and outlier-dominated S14–S15 but worsens fairness on borderline-dominated S13, consistent with the motivating analysis (Appendix I.4).

Table 1: Mean ablation ranks over 19 imbalance settings. Lower is better. Parentheses show method-specific within-metric ranks.

Ablations	Accuracy		Fairness		Overall
	BA	Kappa	SPD	DI	GM_4
WELL w/o ℓ_t^{imb}	4.947 (5)	4.263 (4)	2.632 (3)	2.579 (3)	2.737 (3)
WELL w/o ℓ_t^{dist}	3.211 (3)	3.105 (3)	3.316 (4)	3.474 (4)	3.737 (4)
WELL w/o ℓ_t^{dec}	2.737 (2)	2.789 (2)	4.895 (5)	4.895 (5)	5.000 (5)
WELL w/o ℓ_t^{fair}	**1.105** (1)	**1.000** (1)	5.895 (6)	5.842 (6)	5.211 (6)
WELL w/o d_t	4.947 (6)	5.211 (6)	2.421 (2)	2.421 (2)	2.474 (2)
WELL	4.053 (4)	4.632 (5)	**1.842** (1)	**1.789** (1)	**1.842** (1)

Difficulty-proxy validation. We validate the continuous neighbor-label proxy against the widely used S/B/R/O taxonomy, used only for evaluation. Type-mean scores follow Safe $<$ Borderline $<$ Rare $<$ Outlier across 60 streams (Table 2). This supports sensitivity to its difficulty spectrum, not exact type recovery.

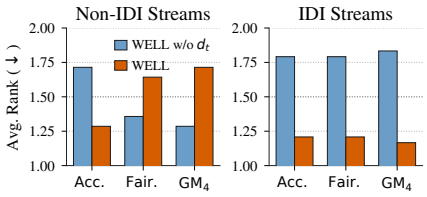


Figure 7: Difficulty-moderation ablation. Bars compare WELL and WELL w/o d_t under non-IDI and IDI streams.

Table 2: kNN proxy validation. "Order" counts seeds whose type means satisfy Safe $<$ Borderline $<$ Rare $<$ Outlier. ✓: ordering holds in all ten seeds.

Composition (S/B/R/O)	Safe	Borderline	Rare	Outlier	Order
0.25 / 0.25 / 0.25 / 0.25	0.112	0.227	0.605	0.608	10/10 ✓
0.70 / 0.10 / 0.10 / 0.10	0.047	0.161	0.749	0.755	10/10 ✓
0.10 / 0.70 / 0.10 / 0.10	0.108	0.169	0.715	0.721	10/10 ✓
0.10 / 0.10 / 0.70 / 0.10	0.265	0.382	0.519	0.520	10/10 ✓
0.10 / 0.10 / 0.10 / 0.70	0.266	0.383	0.517	0.519	10/10 ✓
0.40 / 0.20 / 0.20 / 0.20	0.078	0.198	0.646	0.650	10/10 ✓

Parameter sensitivity. We vary each important coefficient β, γ_{dist}, and γ_{dec} over $\{0, 0.6, 0.8, 1.0, 1.2, 1.4, 2.0, 3.0, 4.0, 5.0\}$ while fixing the others at their default on S0–S18 and denser steps near 1. Figure 8 shows stable mean GM_4 around the unit default. Zeroing any coefficient worsens GM_4, especially γ_{dec}, consistent with the component ablation. Appendix I.1 reports more results.

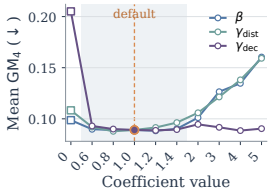


Figure 8: Coefficient sensitivity.

Additional validation. Our experiments demonstrate advantage of continually updating WELL's instance weights over batch and periodic-update alternatives (Appendix I.1.2). Theoretical complexity and runtime analyses further demonstrate its online efficiency, showing bounded per-instance complexity and practical runtime while retaining the strongest overall ranks (Appendix I.1.1).

9 Conclusion and Limitations

Conclusion. Controlled analysis distinguishes CI/SI/IDI fairness effects and intervention limitations. CI and SI affect prediction rates and error disparities differently, while rare- and outlier-dominated streams exhibit stronger fairness degradation under IDI. Model predictions can further amplify disparities already present in stream labels. These findings motivate WELL's cell-level correction and within-cell difficulty moderation, supported by ablations and conditional theory. Across synthetic and real-world streams, WELL outperforms state-of-the-art streaming methods in the overall accuracy–fairness trade-off. These observations emphasize evaluating fairness across imbalance sources and accounting for both evidence availability and reliability during learning.

Limitations and future work. Our stationary analysis does not establish robustness to drift. Low neighbor agreement may reflect informative rare structure rather than noise, and aggregate gains need not protect every subgroup. Future work should examine type-specific errors, estimation uncertainty, and deployment-specific accuracy requirements.

Ethics Statement

This work studies statistical fairness in online classification using controlled synthetic streams and existing benchmark data, including sensitive attributes. Improvements in SPD or DI do not establish that a system is equitable, legally compliant, or suitable for consequential decisions. Historical labels may encode social bias, and aggregate metrics can conceal harms to small subgroups. In particular, low local label agreement must not be interpreted as an individual's lack of merit or eligibility. Any deployment would require application-specific assessment of accuracy costs, subgroup errors, privacy safeguards, and meaningful human oversight. The intended use is methodological evaluation and diagnosis, not automated authorization of high-stakes decisions.

Reproducibility Statement

Algorithm 1 and Section 7 specify the online weighting framework, with experimental settings in Appendix H and the implementation in the released code. Appendix B.2 states the theoretical assumptions and provides full proofs and learner-disagreement bounds. Appendices C, D, and E describe the metrics, stream configurations, and evaluation protocol. Appendix I reports per-setting comparisons and component, sensitivity, difficulty-proxy, and efficiency analyses so that aggregate conclusions can be checked against individual metrics.

AI Use Statement

In this work, we used generative AI tools to aid or polish writing. The authors are responsible for verifying AI-assisted material and for the final content, claims, and interpretations.

References

Census-Income (KDD). UCI Machine Learning Repository, 2000. DOI: https://doi.org/10.24432/C5N30T.

Gabriel Aguiar, Bartosz Krawczyk, and Alberto Cano. A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. *Machine learning*, 113(7):4165–4243, 2024.

Daniel Nowak Assis, Jean Paul Barddal, and Fabrício Enembreck. Just change on change: Adaptive splitting time for decision trees in data stream classification. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, SAC '24, pp. 351–357, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400702433. doi: 10.1145/3605098.3635899. URL https://doi.org/10.1145/3605098.3635899.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Alessio Bernardo and Emanuele Della Valle. SMOTE-OB: Combining smote and online bagging for continuous rebalancing of evolving data streams. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 5033–5042. IEEE, 2021a.

Alessio Bernardo and Emanuele Della Valle. VFC-SMOTE: very fast continuous synthetic minority oversampling for evolving data streams. *Data Mining and Knowledge Discovery*, 35(6):2679–2713, 2021b.

Alessio Bernardo, Emanuele Della Valle, and Albert Bifet. Incremental rebalancing learning on evolving data streams. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pp. 844–850. IEEE, 2020.

Albert Bifet, Geoff Holmes, Richard Kirkby, and Bernhard Pfahringer. MOA: Massive online analysis. *Journal of Machine Learning Research*, 11(52):1601–1604, 2010a. URL http://jmlr.org/papers/v11/bifet10a.html.

Albert Bifet, Geoff Holmes, and Bernhard Pfahringer. Leveraging bagging for evolving data streams. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 135–150. Springer, 2010b.

Dariusz Brzezinski, Leandro Minku, Tomasz Pewinski, Jerzy Stefanowski, and Artur Szumaczuk. The impact of data difficulty factors on classification of imbalanced and concept drifting data streams. *Knowledge and Information Systems*, 63(6):1429–1469, 2021.

Alberto Cano and Bartosz Krawczyk. Kappa updated ensemble for drifting data stream mining. *Machine Learning*, 109(1):175–218, 2020.

Alberto Cano and Bartosz Krawczyk. ROSE: robust online self-adjusting ensemble for continual learning on imbalanced drifting data streams. *Machine Learning*, 111(7):2561–2599, 2022.

Pedro Domingos and Geoff Hulten. Mining high-speed data streams. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 71–80, 2000.

Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61:863–905, 2018.

Luis Eduardo Boiko Ferreira, Heitor Murilo Gomes, Albert Bifet, and Luiz S Oliveira. Adaptive random forests with resampling for imbalanced data streams. In *2019 International Joint Conference on Neural Networks*, pp. 1–6. IEEE, 2019.

Xinyi Gao, Dongting Xie, Yihang Zhang, Zhengren Wang, Chong Chen, Conghui He, Hongzhi Yin, and Wentao Zhang. A comprehensive survey on imbalanced data learning. *Frontiers of Computer Science*, 20(11):2011622, 2026.

Subhodeep Ghosh, Zhihui Du, Angela Bonifati, Manish Kumar, David Bader, and Senjuti Basu Roy. Continuous fairness on data streams. *arXiv preprint arXiv:2601.08976*, 2026.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. *Advances in Neural Information Processing Systems*, 31, 2018.

Heitor M Gomes, Albert Bifet, Jesse Read, Jean Paul Barddal, Fabrício Enembreck, Bernhard Pfharinger, Geoff Holmes, and Talel Abdessalem. Adaptive random forests for evolving data stream classification. *Machine Learning*, 106(9):1469–1495, 2017.

Nuwan Gunasekara, Bernhard Pfahringer, Heitor Gomes, and Albert Bifet. Gradient boosted trees for evolving data streams. *Machine Learning*, 113(5):3325–3352, 2024.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

Geoff Hulten, Laurie Spencer, and Pedro Domingos. Mining time-changing data streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pp. 97–106, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 158113391X. doi: 10.1145/502512.502529. URL https://doi.org/10.1145/502512.502529.

Vasileios Iosifidis and Eirini Ntoutsi. FABBOO-online fairness-aware learning under class imbalance. In *International Conference on Discovery Science*, pp. 159–174. Springer, 2020.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Ron Kohavi. Census Income. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5GP7S.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

Kathrin Lammers, Valerie Vaquet, and Barbara Hammer. Continuous fair SMOTE–fairness-aware stream learning from imbalanced data. In *International Conference on Artificial Neural Networks*, pp. 336–348. Springer, 2025.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. Data and analysis for "how we analyzed the COMPAS recidivism algorithm", 2016. URL `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`.

Junghyun Lee, Hanseul Cho, Se-Young Yun, and Chulhee Yun. Fair streaming principal component analysis: Statistical and algorithmic viewpoint. *Advances in Neural Information Processing Systems*, 36:5126–5167, 2023.

Michelle Seng Ah Lee, Kirtan Padh, David Watson, Niki Kilbertus, and Jatinder Singh. Fairness under uncertainty in sequential decisions. *arXiv preprint arXiv:2604.21711*, 2026.

Peizhao Li and Hongfu Liu. Achieving fairness at no utility cost via data reweighing with influence. In *International conference on machine learning*, pp. 12917–12930. PMLR, 2022.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Weike Liu, Hang Zhang, Zhaoyun Ding, Qingbao Liu, and Cheng Zhu. A comprehensive active learning method for multiclass imbalanced data streams with concept drift. *Knowledge-Based Systems*, 215:106778, 2021.

Lucas Loezer, Fabrício Enembreck, Jean Paul Barddal, and Alceu de Souza Britto Jr. Cost-sensitive learning for imbalanced data streams. In *Proceedings of the 35th annual ACM Symposium on Applied Computing*, pp. 498–504, 2020.

ROBERT LYON, John Brooke, Joshua Knowles, and Benjamin Stappers. Hellinger distance trees for imbalanced streams. *Proceedings - International Conference on Pattern Recognition*, 2014 22nd International Conference on Pattern Recognition, December 2014. ISSN 1051-4651. doi: 10.1109/ICPR.2014.344.

Krystyna Napierala and Jerzy Stefanowski. Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, 46(3): 563–597, 2016.

Nikunj C. Oza and Stuart J. Russell. Online bagging and boosting. In Thomas S. Richardson and Tommi S. Jaakkola (eds.), *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, volume R3 of *Proceedings of Machine Learning Research*, pp. 229–236. PMLR, 04–07 Jan 2001.

Diem Pham, Binh Tran, Su Nguyen, Damminda Alahakoon, and Mengjie Zhang. Fairness optimisation with multi-objective swarms for explainable classifiers on data streams. *Complex & Intelligent Systems*, 10(4):4741–4754, 2024.

Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. An instance level analysis of data complexity. *Machine Learning*, 95(2):225–256, 2014.

Shivashankar Subramanian, Afshin Rahimi, Timothy Baldwin, Trevor Cohn, and Lea Frermann. Fairness-aware class imbalanced learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2045–2051, 2021.

Davoud Ataee Tarzanagh, Bojian Hou, Boning Tong, Qi Long, and Li Shen. Fairness-aware class imbalanced learning on multiple subgroups. In *Uncertainty in Artificial Intelligence*, pp. 2123–2133. PMLR, 2023.

Shuo Wang, Leandro L Minku, Xin Yao, et al. Dealing with multiple classes in online class imbalance learning. In *IJCAI*, pp. 2118–2124, 2016.

Shuo Wang, Leandro L Minku, and Xin Yao. A systematic study of online class imbalance learning with concept drift. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10):4802–4821, 2018.

Qingquan Zhang, Jialin Liu, Zeqi Zhang, Junyi Wen, Bifei Mao, and Xin Yao. Mitigating unfairness via evolutionary multiobjective ensemble learning. *IEEE Transactions on Evolutionary Computation*, 27(4):848–862, 2023. doi: 10.1109/TEVC.2022.3209544.

Qingquan Zhang, Jialin Liu, and Xin Yao. Fairness-aware multiobjective evolutionary learning. *IEEE Transactions on Evolutionary Computation*, 29(6):2372–2385, 2025. doi: 10.1109/TEVC.2024.3430824.

Wenbin Zhang and Eirini Ntoutsi. FAHT: an adaptive fairness-aware decision tree classifier. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 1480–1486, 2019.

APPENDIX CONTENTS

A Broader Impacts **16**

B Proofs **16**

B.1 Profile-based Fairness Evaluation 16

B.2 Theoretical justification of WELL 17

C Evaluation Metrics **24**

C.1 Accuracy metrics 24

C.2 Fairness metrics 25

D Datasets and Stream Construction **26**

D.1 Overview 26

D.2 Base synthetic stream generators 26

D.3 Controlling CI, SI, and data-side fairness 27

D.4 Diagnostic imbalance grids 28

D.5 IDI diagnostic grids 29

D.6 Representative multi-source imbalance settings 29

D.7 Real-world streams 30

D.8 Synthetic-to-real Profile and Behavior Consistency 30

E Experimental Protocol and Compared Methods **31**

E.1 Compute Resources 31

E.2 Shared Prequential Evaluation Protocol 31

E.3 Statistical Analysis for the Motivating Analysis 31

E.4 Algorithms Used in the Motivating Analysis 32

E.5 Fairness-aware Methods Used in RQ1 33

E.6 Methods and Ablations Used in RQ2 33

E.7 Ranking and GM_4 Computation 33

F Additional Results for the Motivating Analysis **34**

F.1 Overview 34

F.2 Preliminaries 34

F.3 Detailed evidence for Finding 1: CI exposes algorithmic unfairness through group-wise error disparities. 35

F.4 Detailed evidence for Finding 2: SI broadly degrades group-based fairness in algorithmic decisions. 40

F.5 Detailed evidence for Finding 3: Imbalance amplifies algorithmic unfairness inherited from biased data streams. 44

F.6 Detailed evidence for Finding 4: IDI severity amplifies fairness degradation, especially under rare and outlier dominance. 45

F.7 Additional Finding: Imbalance sources alter agreement among algorithmic performance metrics. 54

G Additional Results for RQ1 **54**

G.1 Detailed evidence for Findings 5–7 54

G.2 Additional Finding: Existing methods only partially reduce inherited data-side disparity. 56

H WELL Implementation Details **57**

I Additional Results for RQ2 **58**

I.1 Robustness and Additional Mechanism Checks 58

I.2 Full Comparison on the 19 Synthetic Settings 62

I.3 Full Comparison on Five Real-world Streams 62

I.4 Detailed Ablation on 19 Synthetic Settings 63

I.5 Detailed Ablation on Five Real-world Streams 65

A Broader Impacts

This work studies fair online learning under imbalanced data streams, where models are updated sequentially and may repeatedly affect decision outcomes over time. By analyzing how class imbalance, sensitive-group imbalance, and instance-level difficulty imbalance influence algorithmic fairness, we provide a systematic perspective for understanding fairness degradation in streaming settings. We expect this analysis to help researchers and practitioners diagnose when and why online learners become unfair under different imbalance sources.

Our proposed WELL framework offers a training-time approach to mitigating such effects by reweighting incoming instances during online updates. Instead of treating fairness as a post-processing correction, WELL connects fairness-aware learning with online distributional statistics, group-level imbalance modeling, and instance-level difficulty moderation. We anticipate that this perspective can motivate future studies on fair streaming learning, including drift-aware fairness monitoring, more robust difficulty estimation, and fairness-aware online learning under broader definitions of fairness.

B Proofs

B.1 Profile-based Fairness Evaluation

Some rate-, ratio-, or conditional-rate fairness metrics may be undefined on early prefixes because the empirical denominator of a required conditioning event is zero. Let

$$\mathcal{V}_{t,w}^{\phi} = \left\{\tau \in \mathcal{V}_{t,w} : \phi(\mathcal{R}_{\tau}^{\mathcal{F}}(t,w)) \text{ is well-defined}\right\}.$$

Without smoothing, profile coordinates are evaluated over $\mathcal{V}_{t,w}^{\phi}$. With standard smoothing, all prefixes are valid and $\mathcal{V}_{t,w}^{\phi} = \mathcal{V}_{t,w}$. The theoretical statements below are understood for windows whose endpoint prefix is valid, i.e., $t \in \mathcal{V}_{t,w}^{\phi}$, or equivalently under smoothing.

For scalar comparison, write $\mathcal{V}_{t,w}^{\phi} = \{\tau_1 < \cdots < \tau_m\}$. We use an exponentially decayed weight vector

$$\lambda_i = \frac{\eta^{m-i}}{\sum_{j=1}^{m} \eta^{m-j}}, \qquad i = 1, \ldots, m,$$

where $0 < \eta \leq 1$. More recent prefixes therefore receive larger weights, and $\eta = 1$ recovers uniform averaging. When smoothing is used, $m = w$, which matches the full-window profile definition in the main text.

Proof of Proposition 1. The projection statement follows directly from the definition of the profile. Since the endpoint prefix is valid and t is the last prefix in the window,

$$E_{\phi}^{\mathcal{F}}(t,w) = \phi(\mathcal{R}_{t}^{\mathcal{F}}(t,w)) = \pi_{\text{last}}\left(\mathbf{P}_{\phi}^{\mathcal{F}}(t,w)\right).$$

We next show that this projection generally discards information. For many standard rate-based fairness metrics, the endpoint score is a function only of the terminal empirical contingency table of the window. In contrast, the profile is a function of the entire prefix path of empirical contingency tables. Thus, endpoint-only evaluation identifies all prequential traces with the same terminal table, whereas profile-based evaluation can distinguish traces that arrive at the same terminal table through different intermediate paths.

We instantiate this non-injectivity using demographic parity difference. For a prefix ending at τ, define the empirical contingency table

$$C_{\tau}(s,\hat{y}) = \sum_{u=t-w+1}^{\tau} \mathbf{1}(s_u = s, \hat{y}_u = \hat{y}).$$

The demographic parity difference at prefix τ is

$$\phi_{\text{DP}}(\mathcal{R}_{\tau}^{\mathcal{F}}(t,w)) = \left| \frac{C_{\tau}(0,1)}{C_{\tau}(0,0) + C_{\tau}(0,1)} - \frac{C_{\tau}(1,1)}{C_{\tau}(1,0) + C_{\tau}(1,1)} \right|,$$

whenever both denominators are nonzero. Hence the endpoint score depends only on the terminal table C_t, while the profile depends on the sequence of prefix tables $(C_\tau)_{\tau \in \mathcal{V}^{\phi}_{t,w}}$.

Consider two traces of $(S, \hat{Y})$:

$$T = \big((0,1),(1,0),(0,0),(1,1)\big), \qquad T' = \big((0,1),(1,1),(0,0),(1,0)\big).$$

The label Y is irrelevant for demographic parity difference and can be chosen arbitrarily. Both traces have the same terminal contingency table: each sensitive group has one positive and one negative prediction. Therefore, their endpoint demographic parity differences are identical:

$$E^{\mathcal{F}}_{\phi_{\mathrm{DP}}}(T) = E^{\mathcal{F}}_{\phi_{\mathrm{DP}}}(T') = 0.$$

However, the two traces have different prefix paths. For T, the valid-prefix demographic parity differences are

$$|1-0| = 1, \qquad \left|\frac{1}{2} - 0\right| = \frac{1}{2}, \qquad \left|\frac{1}{2} - \frac{1}{2}\right| = 0,$$

and therefore

$$\mathbf{P}^{\mathcal{F}}_{\phi_{\mathrm{DP}}}(T) = (1, \tfrac{1}{2}, 0).$$

For T', the valid-prefix demographic parity differences are

$$|1-1| = 0, \qquad \left|\frac{1}{2} - 1\right| = \frac{1}{2}, \qquad \left|\frac{1}{2} - \frac{1}{2}\right| = 0,$$

and therefore

$$\mathbf{P}^{\mathcal{F}}_{\phi_{\mathrm{DP}}}(T') = (0, \tfrac{1}{2}, 0).$$

Thus,

$$\mathbf{P}^{\mathcal{F}}_{\phi_{\mathrm{DP}}}(T) \neq \mathbf{P}^{\mathcal{F}}_{\phi_{\mathrm{DP}}}(T'), \qquad E^{\mathcal{F}}_{\phi_{\mathrm{DP}}}(T) = E^{\mathcal{F}}_{\phi_{\mathrm{DP}}}(T').$$

This establishes that endpoint-only evaluation is a many-to-one projection of the prefix-path profile and can discard transient fairness behavior within the window. □

B.2 Theoretical justification of WELL

This section proves Lemma 1 and Theorem 1, then strengthens their conclusions in three directions. We derive the fairness-correction directions from WELL's actual log-weight terms, characterize joint SPD/DI improvement beyond the no-crossing case, and refine the accuracy bound using posterior uncertainty. A separate result establishes within-cell influence control by difficulty moderation. The analysis fixes one pre-update history and characterizes weighted population responses, with an explicit extension to approximate learners.

Matched comparison and conditioning. Fix the pre-update history $\mathcal{I}_t$. Conditional on $\mathcal{I}_t$, all discounted statistics, the current decision-disparity multiplier a_t^{dec}, and the resulting cell-level log terms are fixed. We compare the full WELL rule with a matched rule that retains the same group-imbalance and difficulty factors but removes only the distributional-association and decision-disparity terms. This comparator is useful because it changes neither the observed stream state nor the treatment of locally inconsistent instances. Consequently, any difference between the two weighted update distributions is attributable to the two fairness corrections.

For a label–sensitive cell $g = (y, s)$, define

$$\Gamma_t(g) = \gamma_{\mathrm{dist}} \ell_t^{\mathrm{dist}}(g) + \gamma_{\mathrm{dec}} a_t^{\mathrm{dec}} \ell_t^{\mathrm{dec}}(g), \tag{9}$$

$$L_t^{\mathrm{matched}}(g) = \beta \ell_t^{\mathrm{imb}}(g), \qquad L_t^{\mathrm{full}}(g) = \beta \ell_t^{\mathrm{imb}}(g) + \Gamma_t(g). \tag{10}$$

The difficulty-moderated probability mass of cell g is

$$q_t^{\chi}(g) = \mathbb{E}[\mathbf{1}\{(Y,S) = g\}\chi(d_t) \mid \mathcal{I}_t]\,. \tag{11}$$

This quantity incorporates both the prevalence of the cell and the average influence retained after instance-level difficulty moderation. The positivity condition in Lemma 1 ensures that ratios involving $q_t^{\chi}(g)$ are well defined.

To retain the feature- and label-dependence of difficulty moderation in the pointwise analysis, let $Z_t^\chi = \mathbb{E}[\chi(d_t) \mid \mathcal{I}_t] = \sum_{g\in\mathcal{G}} q_t^\chi(g)$ and define

$$Q_t^\chi(A) = \frac{\mathbb{E}[\mathbf{1}\{(X,Y,S)\in A\}\chi(d_t) \mid \mathcal{I}_t]}{Z_t^\chi}. \tag{12}$$

Equation equation 12 does not assume that $\chi(d_t)$ is constant within a (Y,S) cell; its dependence on the current features and label is retained in Q_t^χ.

Proof of Lemma 1. The unnormalized weighted masses assigned to cell g are

$$U_t^r(g) = q_t^\chi(g)\exp(L_t^r(g))\,, \qquad r\in\{\text{matched},\text{full}\}. \tag{13}$$

Let $Z_t^r = \sum_{h\in\mathcal{G}} U_t^r(h)$ and $\mathcal{W}_t^r(g) = U_t^r(g)/Z_t^r$. For any two positive-mass cells g and g', the common normalizer cancels:

$$\frac{\mathcal{W}_t^r(g)}{\mathcal{W}_t^r(g')} = \frac{U_t^r(g)}{U_t^r(g')}. \tag{14}$$

Substitution of Eq. equation 10 gives

$$\frac{\mathcal{W}_t^{\text{matched}}(g)}{\mathcal{W}_t^{\text{matched}}(g')} = \frac{q_t^\chi(g)}{q_t^\chi(g')} e^{\beta[\ell_t^{\text{imb}}(g)-\ell_t^{\text{imb}}(g')]}, \tag{15}$$

$$\frac{\mathcal{W}_t^{\text{full}}(g)}{\mathcal{W}_t^{\text{full}}(g')} = \frac{q_t^\chi(g)}{q_t^\chi(g')} e^{\beta[\ell_t^{\text{imb}}(g)-\ell_t^{\text{imb}}(g')]+\Gamma_t(g)-\Gamma_t(g')}. \tag{16}$$

Equivalently, the change in relative log influence is

$$\log\frac{\mathcal{W}_t^{\text{full}}(g)/\mathcal{W}_t^{\text{full}}(g')}{\mathcal{W}_t^{\text{matched}}(g)/\mathcal{W}_t^{\text{matched}}(g')} = \Gamma_t(g)-\Gamma_t(g'). \tag{17}$$

Equation equation 17 makes the coefficient roles explicit. The coefficient β affects the relative influence of scarce cells in both rules. It cancels only from the incremental full-versus-matched comparison because the group-imbalance factor is deliberately held fixed. In contrast, γ_{dist} and γ_{dec} directly scale the additional relative shift through Γ_t. The latter contribution is also graded by a_t^{dec}, so prediction-side correction increases continuously with the disparity present in the current prequential history.

It remains to connect the cell factors with pointwise decisions. Define the original conditional posterior and the conditional expected difficulty factors by

$$\eta_{t,s}(x) = P(Y=1 \mid X=x, S=s, \mathcal{I}_t), \qquad \bar\chi_{t,ys}(x) = \mathbb{E}[\chi(d_t) \mid X=x, Y=y, S=s, \mathcal{I}_t]. \tag{18}$$

The posterior under the common difficulty-moderated law is therefore

$$\eta_{t,s}^\chi(x) = \frac{\bar\chi_{t,1s}(x)\eta_{t,s}(x)}{\bar\chi_{t,1s}(x)\eta_{t,s}(x) + \bar\chi_{t,0s}(x)[1-\eta_{t,s}(x)]}, \tag{19}$$

whenever the denominator is positive. Thus, the feature- and label-dependence of $\chi(d_t)$ changes the common posterior rather than disappearing from the analysis. For $r\in\{\text{matched},\text{full}\}$, set $w_{ys}^r = \exp(L_t^r(y,s))$. Because difficulty moderation is absorbed into Q_t^χ, the pointwise r-rule binary log-loss risk is

$$\mathcal{R}_{t,s}^r(p;x) = -w_{1s}^r\eta_{t,s}^\chi(x)\log p - w_{0s}^r[1-\eta_{t,s}^\chi(x)]\log(1-p). \tag{20}$$

Its first two derivatives are

$$\frac{\partial\mathcal{R}_{t,s}^r}{\partial p} = -\frac{w_{1s}^r\eta_{t,s}^\chi(x)}{p} + \frac{w_{0s}^r[1-\eta_{t,s}^\chi(x)]}{1-p}, \tag{21}$$

$$\frac{\partial^2\mathcal{R}_{t,s}^r}{\partial p^2} = \frac{w_{1s}^r\eta_{t,s}^\chi(x)}{p^2} + \frac{w_{0s}^r[1-\eta_{t,s}^\chi(x)]}{(1-p)^2} > 0. \tag{22}$$

For $0<\eta_{t,s}^\chi(x)<1$, the risk is strictly convex, and solving the first-order condition yields the unique interior minimizer

$$p_{t,s}^{r,*}(x) = \frac{w_{1s}^r\eta_{t,s}^\chi(x)}{w_{1s}^r\eta_{t,s}^\chi(x) + w_{0s}^r[1-\eta_{t,s}^\chi(x)]}. \tag{23}$$

For $\eta^{\chi}_{t,s}(x) \in \{0,1\}$, the same expression gives the corresponding boundary minimizer by continuous extension. Using the usual $1/2$ cutoff,

$$p^{r,*}_{t,s}(x) \geq \frac{1}{2} \quad \Longleftrightarrow \quad \eta^{\chi}_{t,s}(x) \geq \tau^{r}_{t,s} := \frac{w^{r}_{0s}}{w^{r}_{0s} + w^{r}_{1s}}. \tag{24}$$

$$\log \frac{w^{\text{full}}_{1s}/w^{\text{full}}_{0s}}{w^{\text{matched}}_{1s}/w^{\text{matched}}_{0s}} = \Gamma_t(1,s) - \Gamma_t(0,s). \tag{25}$$

Therefore, the two rules act on the same difficulty-moderated posterior, and their threshold difference is attributable only to the fairness corrections. This establishes both the cell-ratio and threshold responses stated in the lemma without suppressing the within-cell dependence of $\chi(d_t)$. □

When the corrections improve parity. The response direction can be derived from the statistics in Eqs. equation 3–equation 4. Write $\widehat{\mu} = \hat{p}_t(Y = 1)$ and $\widehat{\mu}_s = \hat{p}_t(Y = 1 \mid S = s)$ for the empirical label rates, and $\widehat{r} = \hat{p}^{\text{pred}}_t(\hat{Y} = 1)$ and $\widehat{r}_s = \hat{p}^{\text{pred}}_t(\hat{Y} = 1 \mid S = s)$ for the historical prediction rates. These estimates are fixed by the history and need not equal the population response rates r_H, r_L. Assume positive empirical cells and nonnegative outer coefficients, and write $\text{logit}(u) = \log[u/(1-u)]$.

Proposition 2 (Statistics-induced correction direction)**.** *Let $D_s = \Gamma_t(1,s) - \Gamma_t(0,s)$ and $o_s = w^{\text{matched}}_{1s}/w^{\text{matched}}_{0s} > 0$. Then*

$$\begin{aligned} D_s = \gamma_{\text{dist}} \big[\text{logit}(\widehat{\mu}) - \text{logit}(\widehat{\mu}_s)\big] \\ + \gamma_{\text{dec}} a^{\text{dec}}_t \big[\text{logit}(\widehat{r}) - \text{logit}(\widehat{r}_s)\big]. \end{aligned} \tag{26}$$

The corresponding thresholds are

$$\tau^{\text{matched}}_{t,s} = \frac{1}{1+o_s}, \qquad \tau^{\text{full}}_{t,s} = \frac{1}{1+o_s e^{D_s}}. \tag{27}$$

If the empirical label and prediction rates satisfy $\widehat{\mu}_H \geq \widehat{\mu} \geq \widehat{\mu}_L$ and $\widehat{r}_H \geq \widehat{r} \geq \widehat{r}_L$, where H, L index the matched response's higher- and lower-positive-rate groups, then $D_H \leq 0 \leq D_L$. The full-rule threshold is therefore no smaller in H and no larger in L than the matched threshold. No equality of sensitive-group prevalences is required.

Proof. The cell-independent centering constants cancel when subtracting the log terms for $(1,s)$ and $(0,s)$. For the distributional term,

$$\ell^{\text{dist}}_t(1,s) - \ell^{\text{dist}}_t(0,s) = \log \frac{\hat{p}_t(Y=1)\hat{p}_t(0,s)}{\hat{p}_t(Y=0)\hat{p}_t(1,s)} = \text{logit}(\widehat{\mu}) - \text{logit}(\widehat{\mu}_s).$$

Applying the same calculation to the prediction-side table proves Eq. equation 26. Lemma 1 yields Eq. equation 27. Monotonicity of the logit and nonnegativity of the coefficients establish the signs of D_H, D_L, and the thresholds have the stated directions because $u \mapsto (1 + o_s e^u)^{-1}$ is decreasing. □

Thus the two corrections act as negative feedback on group-specific label and prediction odds when these statistics are aligned with the matched response. If the two sources disagree, Eq. equation 26 determines whether their combined effect still satisfies $D_H \leq 0 \leq D_L$. Alignment is a sufficient condition, not a consequence of stationarity: historical predictions need not have the same group ordering as the matched population response. The coefficient β and the difficulty term determine that response and its thresholds, even though they are held fixed in the incremental comparison.

Parity response. Although the response classifiers are induced by weighted-risk minimization under Q^{χ}_t, all rates, switch masses, and accuracy quantities below are measured under the original conditional population law $P_t = P(\cdot \mid \mathcal{I}_t)$. The theorem studies the case in which the induced threshold shifts act in the parity-correcting direction. Let H and L denote the higher- and lower-positive-rate groups under the matched response, with $r_H \geq r_L > 0$. Relative to this response, suppose a group-conditional mass m_H changes from $1 \to 0$ in H, while a mass m_L changes from $0 \to 1$ in L. The full-rule positive rates are then

$$r^{\text{full}}_H = r_H - m_H, \qquad r^{\text{full}}_L = r_L + m_L. \tag{28}$$

The no-crossing assumption $r_H - m_H \geq r_L + m_L > 0$ preserves the identities of the higher- and lower-rate groups. It also fixes the relevant branches of the absolute SPD and minimum-symmetric DI definitions.

Proof of Theorem 1. Under the preceding ordering,

$$\mathrm{SPD}_{\mathrm{matched}} = r_H - r_L, \qquad \mathrm{SPD}_{\mathrm{full}} = (r_H - m_H) - (r_L + m_L), \tag{29}$$

$$\mathrm{DI}_{\mathrm{matched}} = 1 - \frac{r_L}{r_H}, \qquad \mathrm{DI}_{\mathrm{full}} = 1 - \frac{r_L + m_L}{r_H - m_H}. \tag{30}$$

Subtracting the full-rule values from the matched-rule values gives

$$\mathrm{SPD}_{\mathrm{matched}} - \mathrm{SPD}_{\mathrm{full}} = m_H + m_L, \tag{31}$$

$$\mathrm{DI}_{\mathrm{matched}} - \mathrm{DI}_{\mathrm{full}} = \frac{r_H m_L + r_L m_H}{r_H(r_H - m_H)}. \tag{32}$$

The denominator in Eq. equation 32 is positive, and the numerator is nonnegative. Hence neither SPD nor DI worsens under the stated response conditions, and both improve strictly whenever $m_H + m_L > 0$. The SPD improvement equals the total parity-directed rate shift exactly. The DI improvement additionally depends on the original group rates because DI is a relative rather than an absolute disparity measure.

The same switch sets yield an exact accuracy decomposition. Define

$$a_H = P(Y = 1, \mathrm{switch} \mid S = H), \qquad a_L = P(Y = 1, \mathrm{switch} \mid S = L), \tag{33}$$

where "switch" refers to the corresponding decision change in each group. Among the $1 \to 0$ changes in H, mass $m_H - a_H$ corrects negative-label errors and mass a_H introduces positive-label errors. Among the $0 \to 1$ changes in L, mass a_L corrects positive-label errors and mass $m_L - a_L$ introduces negative-label errors. Consequently,

$$\mathrm{Acc}_{\mathrm{full}} - \mathrm{Acc}_{\mathrm{matched}} = P(S = H)[(m_H - a_H) - a_H] + P(S = L)[a_L - (m_L - a_L)] \tag{34}$$

$$= P(S = H)(m_H - 2a_H) + P(S = L)(2a_L - m_L). \tag{35}$$

Let

$$C_t = P(S = H)(m_H - a_H) + P(S = L)a_L, \tag{36}$$

$$E_t = P(S = H)a_H + P(S = L)(m_L - a_L) \tag{37}$$

denote the prevalence-weighted masses of corrected and newly introduced errors. Equation equation 35 is equivalently $\mathrm{Acc}_{\mathrm{full}} - \mathrm{Acc}_{\mathrm{matched}} = C_t - E_t$. Accuracy therefore improves when $C_t > E_t$, remains unchanged when $C_t = E_t$, and decreases only by the excess newly introduced error mass when $C_t < E_t$. Since $C_t \geq 0$ and $E_t \leq P(S = H)m_H + P(S = L)m_L$, we obtain

$$\mathrm{Acc}_{\mathrm{full}} \geq \mathrm{Acc}_{\mathrm{matched}} - P(S = H)m_H - P(S = L)m_L. \tag{38}$$

Thus the theorem does not assume that fairness improvement is cost free. Instead, it bounds the worst possible accuracy loss by the probability mass on which decisions change.

For completeness, let $q_s = P(Y = 1 \mid S = s)$ and consider the intersectional BA used in the experiments. The four switch contributions give

$$\mathrm{BA}_{\mathrm{full}} - \mathrm{BA}_{\mathrm{matched}} = \frac{1}{4}\left[-\frac{a_H}{q_H} + \frac{m_H - a_H}{1 - q_H} + \frac{a_L}{q_L} - \frac{m_L - a_L}{1 - q_L}\right]. \tag{39}$$

The second and third terms are corrections in the $(Y = 0, S = H)$ and $(Y = 1, S = L)$ cells, whereas the first and fourth terms are newly introduced errors. Dropping the two nonnegative correction terms yields the conservative bound

$$\mathrm{BA}_{\mathrm{full}} \geq \mathrm{BA}_{\mathrm{matched}} - \frac{1}{4}\left(\frac{a_H}{q_H} + \frac{m_L - a_L}{1 - q_L}\right). \tag{40}$$

As with ordinary accuracy, BA is nondecreasing whenever the normalized mass of corrected errors is at least the normalized mass of newly introduced errors. This proves all statements in Theorem 1. □

Joint fairness improvement beyond no crossing. Theorem 1 gives a simple sufficient regime in which the group-rate ordering is preserved. The following result characterizes a larger improvement region, including responses that reverse this ordering without increasing either disparity measure. Throughout, B and F denote the matched and full population responses, and probabilities are evaluated under P_t, not the reweighted training law.

Theorem 2 (Exact joint parity-improvement region)**.** *Consider the population responses in Lemma 1, with matched positive rates $r_H > r_L > 0$ and $D_H \le 0 \le D_L$. Define the switching sets*

$$\begin{aligned} A_H &= \{x : \tau^B_{t,H} \le \eta^{\chi}_{t,H}(x) < \tau^F_{t,H}\}, \\ A_L &= \{x : \tau^F_{t,L} \le \eta^{\chi}_{t,L}(x) < \tau^B_{t,L}\}, \qquad m_s = P_t(X \in A_s \mid S = s). \end{aligned} \tag{41}$$

Let $G = r_H - r_L$, $m = m_H + m_L$, $u = r_H - m_H$, $v = r_L + m_L$, and $\pi_s = P_t(S = s)$. Then

$$\begin{gathered} \mathrm{SPD}_F = |G - m|, \qquad \mathrm{DI}_F = \frac{|G - m|}{\max\{u, v\}}, \\ \mathrm{Acc}_F \ge \mathrm{Acc}_B - \pi_H m_H - \pi_L m_L. \end{gathered} \tag{42}$$

Both SPD and DI improve strictly if and only if

$$0 < m < 2G, \qquad r_H m_H + r_L m_L < (r_H + r_L)G. \tag{43}$$

In particular, $0 < m \le G$ recovers the no-crossing regime, but no-crossing is not necessary for simultaneous improvement.

Proof. Proposition 2 and Lemma 1 imply that predictions can change only from 1 to 0 in H and from 0 to 1 in L, on the sets in Eq. equation 41. Their new positive rates are u and v. Since $v \ge r_L > 0$, the symmetric DI definition gives $1 - \min\{u, v\}/\max\{u, v\} = |u - v|/\max\{u, v\}$. This proves the first two identities in Eq. equation 42. The accuracy bound follows from the disagreement mass $\pi_H m_H + \pi_L m_L$, regardless of whether the rates cross.

The matched SPD is G, so strict SPD improvement is equivalent to $|G - m| < G$, or $0 < m < 2G$. When $0 < m \le G$, the full rates do not cross and Eq. equation 32 gives strict DI improvement. In this case the second inequality in Eq. equation 43 is automatic, because

$$r_H m_H + r_L m_L \le r_H m \le r_H G < (r_H + r_L)G.$$

When $m > G$, the rates cross and $\mathrm{DI}_F = 1 - u/v$. Strict DI improvement is then equivalent to

$$\begin{aligned} \frac{u}{v} > \frac{r_L}{r_H} \quad &\iff \quad r_H(r_H - m_H) > r_L(r_L + m_L) \\ &\iff \quad r_H m_H + r_L m_L < (r_H + r_L)G. \end{aligned}$$

Combining the two cases proves the necessary and sufficient joint conditions. □

For example, $r_H = 0.8$, $r_L = 0.4$, $m_H = 0.25$, and $m_L = 0.30$ yield full rates 0.55 and 0.70. Although the ordering reverses, SPD decreases from 0.40 to 0.15 and DI from 0.50 to approximately 0.214. This illustrates the strictly larger region covered by Theorem 2. Conversely, the conditions identify when an excessively large correction loses the joint improvement.

From weight changes to response-size conditions. The switch masses in Theorem 2 depend on the posterior mass between the thresholds. Their size can therefore be bounded from the log-weight shift and the local distribution of $\eta^{\chi}_{t,s}(X)$, rather than imposed independently of WELL's weighting rule.

Corollary 3 (Sufficient conditions from threshold displacement)**.** *Under the assumptions of Theorem 2, suppose the conditional distribution of $\eta^{\chi}_{t,s}(X)$ given $S = s$ under P_t has density bounded by C_s on the interval between the two thresholds, with no atoms at its endpoints. Set*

$$\Delta\tau_s = \frac{o_s|e^{D_s} - 1|}{(1 + o_s)(1 + o_s e^{D_s})}, \qquad b_s = C_s \Delta\tau_s. \tag{44}$$

Then $m_s \le b_s$ and $\Delta\tau_s \le |D_s|/4$. If $m > 0$ and

$$b_H + b_L < 2G, \qquad r_H b_H + r_L b_L < (r_H + r_L)G, \tag{45}$$

both SPD and DI improve strictly, with accuracy loss at most $\pi_H b_H + \pi_L b_L$.

Proof. Subtracting the thresholds in Eq. equation 27 yields $\Delta\tau_s = |\tau^F_{t,s} - \tau^B_{t,s}|$. For $h_s(z) = (1+o_s e^z)^{-1}$, $|h'_s(z)| = o_s e^z/(1+o_s e^z)^2 \le 1/4$, giving the second bound by the mean value theorem. Integrating the conditional density over the switching interval gives $m_s \le C_s \Delta\tau_s$. Substitution into Eq. equation 43 establishes joint fairness improvement, and Eq. equation 42 gives the accuracy bound. □

Here C_s describes a population distribution and is not an additional algorithm parameter. The corollary is a sufficient population condition, not an assumption needed to run WELL. When the posterior distribution has atoms, Theorem 2 still applies using the exact switch probabilities. Together, these results connect the observed label and prediction statistics to the log-weight shift, threshold displacement, and resulting fairness change.

A sharper accuracy bound from posterior uncertainty. The disagreement-mass bound allows every changed decision to be harmful. The next result retains the original label posterior $\eta_{t,s}(x) = P_t(Y = 1 \mid X = x, S = s)$ and quantifies the smaller accuracy cost when changed decisions lie near its uncertainty region. This posterior is distinct from the difficulty-moderated posterior that determines the switching sets.

Proposition 3 (Posterior-sensitive accuracy bound)**.** *For the switching sets in Eq. equation 41,*

$$\begin{aligned}\mathrm{Acc}_F - \mathrm{Acc}_B = \pi_H \mathbb{E}_{P_t}[(1 - 2\eta_{t,H}(X))\mathbf{1}_{A_H}(X) \mid S = H] \\ + \pi_L \mathbb{E}_{P_t}[(2\eta_{t,L}(X) - 1)\mathbf{1}_{A_L}(X) \mid S = L]\,.\end{aligned} \tag{46}$$

If $|2\eta_{t,s}(X) - 1| \le \zeta_s \le 1$ *almost surely on* A_s*, then*

$$\mathrm{Acc}_F \ge \mathrm{Acc}_B - \pi_H \zeta_H m_H - \pi_L \zeta_L m_L. \tag{47}$$

Moreover, accuracy does not decrease if $\eta_{t,H} \le 1/2$ *on* A_H *and* $\eta_{t,L} \ge 1/2$ *on* A_L*, almost surely.*

Proof. At feature value x in group H, a $1 \to 0$ change replaces correctness probability $\eta_{t,H}(x)$ with $1 - \eta_{t,H}(x)$, giving gain $1 - 2\eta_{t,H}(x)$. The corresponding $0 \to 1$ change in L gives gain $2\eta_{t,L}(x) - 1$. Integrating these gains over the switch sets proves Eq. equation 46. Each integrand is bounded below by $-\zeta_s$ on its switch set, proving Eq. equation 47. Under the final stated conditions, both integrands are nonnegative. □

The refined bound is no weaker than the original mass bound and is tighter whenever a group with positive switch mass has $\zeta_s < 1$. Thus simultaneous fairness improvement need not entail an equally large accuracy decrease. The exact cost depends on the true label uncertainty of the affected instances, not only on how many decisions change. The difficulty proxy enters this bound through the switching sets, without being equated with the original label posterior.

Role of the three coefficients. The derivation also clarifies why β, γ_{dist}, and γ_{dec} remain relevant in the theoretical comparison. The coefficient β determines the shared scarcity-aware baseline and can therefore affect the initial rates r_H, r_L and the switch masses m_H, m_L, although it cancels from the incremental ratio in Eq. equation 17. The coefficients γ_{dist} and γ_{dec} control the magnitude of the additional cell-ratio shift directly through Eq. equation 9. Setting either coefficient to zero removes the corresponding contribution. The decision-side contribution further vanishes when $a^{\mathrm{dec}}_t = 0$. Difficulty moderation remains present through q^{χ}_t in both rules, ensuring that the comparison isolates the two fairness corrections without discarding within-cell reliability.

Within-cell influence control. The preceding comparisons retain the same difficulty factor to isolate the fairness corrections. Its distinct role can also be established directly: unlike the three cell-level terms, $\chi(d_t)$ changes the relative influence of instances sharing the same (Y, S) cell.

Proposition 4 (Within-cell difficulty attenuation)**.** *Suppose* $\chi : [0, 1] \to (0, \infty)$ *is bounded and nonincreasing, as in WELL's difficulty mapping. Fix a positive-mass cell* g*. For any measurable event* A*,*

$$\mathcal{W}^{\mathrm{full}}_t(A \mid g) = \frac{\mathbb{E}_{P_t}[\mathbf{1}_A \chi(d_t) \mid g]}{\mathbb{E}_{P_t}[\chi(d_t) \mid g]}. \tag{48}$$

For $0 \le a < b \le 1$, let $A_{\text{low}} = \{d_t \le a\}$ and $A_{\text{high}} = \{d_t \ge b\}$ have positive conditional probability given g. Then

$$\frac{\mathcal{W}_t^{\text{full}}(A_{\text{high}} \mid g)}{\mathcal{W}_t^{\text{full}}(A_{\text{low}} \mid g)} \le \frac{\chi(b)}{\chi(a)} \frac{P_t(A_{\text{high}} \mid g)}{P_t(A_{\text{low}} \mid g)}, \qquad \mathbb{E}_{\mathcal{W}_t^{\text{full}}}[d_t \mid g] \le \mathbb{E}_{P_t}[d_t \mid g]. \tag{49}$$

These inequalities hold for any fixed cell-level coefficients.

Proof. Conditioning the full weighted law on g cancels its cell-constant factor $\exp(L_t^{\text{full}}(g))$ and gives Eq. equation 48. On A_{high}, $\chi(d_t) \le \chi(b)$, while on A_{low}, $\chi(d_t) \ge \chi(a)$. Applying these bounds to the numerator and denominator of the conditional mass ratio proves the first inequality. For the second, let d, d' be independent copies of d_t conditional on g. Monotonicity gives

$$2\,\mathrm{Cov}_{P_t}(d_t, \chi(d_t) \mid g) = \mathbb{E}_{P_t}[(d - d')(\chi(d) - \chi(d')) \mid g] \le 0.$$

Therefore

$$\mathbb{E}_{\mathcal{W}_t^{\text{full}}}[d_t \mid g] - \mathbb{E}_{P_t}[d_t \mid g] = \frac{\mathrm{Cov}_{P_t}(d_t, \chi(d_t) \mid g)}{\mathbb{E}_{P_t}[\chi(d_t) \mid g]} \le 0.$$

□

When $\chi(b) < \chi(a)$, the high-difficulty subset loses relative influence by at least the stated multiplicative factor, even if its cell is upweighted for scarcity or fairness. Cell-level reweighting alone preserves this within-cell ratio. The result formalizes the complementary roles of cell correction and instance-level moderation, while making no identification of proxy difficulty with label noise or irrelevance.

Intrinsic parity–accuracy lower bound. Let $q_s = P(Y = 1 \mid S = s)$, $r_s = P(\hat{Y} = 1 \mid S = s)$, and $e_s = P(\hat{Y} \ne Y \mid S = s)$ for $s \in \{0, 1\}$. Since a prediction can change the positive rate relative to the label rate only on a misclassified instance,

$$|r_s - q_s| \le e_s. \tag{50}$$

The triangle inequality then gives

$$e_0 + e_1 \ge [|q_0 - q_1| - |r_0 - r_1|]_+ . \tag{51}$$

Writing $\pi_s = P(S = s)$, the overall classification error satisfies

$$P(\hat{Y} \ne Y) \ge \min\{\pi_0, \pi_1\}\,[|q_0 - q_1| - \mathrm{SPD}]_+ . \tag{52}$$

Thus, if the groups have different positive-label rates, forcing prediction SPD below that label-rate difference necessarily incurs a nonzero classification-error burden. This lower bound is method independent and complements Theorem 1: the theorem controls the accuracy change associated with WELL's parity-directed response, while Eqs. equation 51–equation 52 describe an intrinsic limit of parity correction itself.

Approximate realization by an online learner. Let f^r be the population response and $\widetilde{f}^r$ its implemented counterpart, for $r \in \{\text{matched}, \text{full}\}$. Define $\delta_s^r = P_t(\widetilde{f}^r(X, s) \ne f^r(X, s) \mid S = s)$ and $E_s = \delta_s^{\text{matched}} + \delta_s^{\text{full}}$. Each positive rate differs from its population value by at most δ_s^r. Applying the triangle inequality to both rules gives

$$\mathrm{SPD}(\widetilde{f}^{\text{matched}}) - \mathrm{SPD}(\widetilde{f}^{\text{full}}) \ge G - |G - m| - E_H - E_L, \tag{53}$$

$$\mathrm{Acc}(\widetilde{f}^{\text{full}}) \ge \mathrm{Acc}(\widetilde{f}^{\text{matched}}) - \sum_{s \in \{H, L\}} P_t(S = s)(m_s + E_s). \tag{54}$$

Here G and m are as in Theorem 2, so the SPD bound also covers rate crossing and reduces to $m - E_H - E_L$ without crossing. For DI, write $\mathrm{DI}(u, v) = |u - v| / \max(u, v)$. If the maximum positive rate is at least $c > 0$ for each ideal and implemented response, then $|\mathrm{DI}(u, v) - \mathrm{DI}(u', v')| \le 2(|u - u'| + |v - v'|)/c$. Applying this bound to both rules gives

$$\mathrm{DI}(\widetilde{f}^{\text{matched}}) - \mathrm{DI}(\widetilde{f}^{\text{full}}) \ge \frac{G}{r_H} - \frac{|G - m|}{\max\{r_H - m_H, r_L + m_L\}} - \frac{2(E_H + E_L)}{c}. \tag{55}$$

Consequently, the population improvement survives approximate realization whenever its magnitude exceeds the corresponding disagreement term. Under Proposition 3, the accuracy bound also sharpens to $\mathrm{Acc}(\widetilde{f}^{\mathrm{full}}) \geq \mathrm{Acc}(\widetilde{f}^{\mathrm{matched}}) - \sum_s \pi_s(\zeta_s m_s + E_s)$. These bounds characterize the approximation error needed to transfer the population results. They do not assert that the implemented online ensemble attains population risk minimizers or that its disagreement is small at every prefix.

Implications for WELL's design. The analysis explains how the weighting rule connects the empirical findings to two complementary forms of influence control. At the cell level, Proposition 2 expresses the fairness corrections through empirical label and prediction odds. When their combined direction aligns with the matched population response, the resulting threshold shifts are directed toward parity without requiring equal group sizes. Theorem 2 identifies the exact correction range that improves both SPD and DI, including limited rate crossing, while Corollary 3 links that range to WELL's log-weight shifts. Thus, the analysis identifies both when the corrections improve fairness and when an excessive correction can lose this joint improvement.

The accuracy analysis concerns the decisions changed by these corrections. Proposition 3 shows that the cost depends on their original label posteriors as well as their probability mass. When these posteriors are close to $1/2$, the bound is tighter than counting every changed decision as an error. Within a cell, Proposition 4 establishes a different property: difficulty moderation reduces the relative influence of high-proxy-difficulty observations, regardless of the cell-level multiplier. The common multiplier cancels within the cell, so scarcity correction and local moderation can operate together. This influence-control property does not equate high proxy difficulty with mislabeled or dispensable instances. The learner-disagreement bounds specify the approximation error under which the conditional population improvements carry over to implemented responses.

C Evaluation Metrics

C.1 Accuracy metrics

We follow the prequential protocol for imbalanced data streams (Aguiar et al., 2024). For each incoming instance, the current model first predicts the label and is then updated after the true label is revealed. The stream is split into consecutive chunks of length $L = 500$. Let $\mathcal{I}_b = \{(b-1)L+1, \ldots, \min(bL, T)\}$ be the b-th chunk, and let B be the total number of chunks. For an accuracy metric M, we first compute its value on each chunk using the test-then-train predictions and then report

$$\overline{M} = \frac{1}{B}\sum_{b=1}^{B} M\left(\{(s_t, y_t, \hat{y}_t) : t \in \mathcal{I}_b\}\right).$$

In the main experiments, we report intersectional balanced accuracy and Cohen's Kappa. Let $S \in \{0,1\}$ be the sensitive attribute and $Y \in \{0,1\}$ be the class label. For chunk $\mathcal{I}_b$, define the class-sensitive cell

$$\mathcal{I}_b(y,s) = \{t \in \mathcal{I}_b : y_t = y,\ s_t = s\}.$$

The accuracy on cell (y,s) is

$$\mathrm{Acc}_b(y,s) = \frac{1}{|\mathcal{I}_b(y,s)|}\sum_{t \in \mathcal{I}_b(y,s)} \mathbf{1}(\hat{y}_t = y_t),$$

whenever $|\mathcal{I}_b(y,s)| > 0$. The chunk-level balanced accuracy is then defined as the average accuracy over valid class-sensitive cells:

$$\mathrm{BA}_b = \frac{1}{|\mathcal{C}_b|}\sum_{(y,s)\in\mathcal{C}_b} \mathrm{Acc}_b(y,s), \qquad \mathcal{C}_b = \{(y,s) : |\mathcal{I}_b(y,s)| > 0\}.$$

When all four binary cells are nonempty, this reduces to

$$\mathrm{BA}_b = \frac{1}{4}\sum_{y\in\{0,1\}}\sum_{s\in\{0,1\}} P(\hat{Y} = Y \mid Y = y, S = s).$$

This metric balances accuracy across the joint class-sensitive cells, rather than only across class labels, and is therefore sensitive to both class imbalance and sensitive-group imbalance.

Cohen's Kappa is computed on each chunk as

$$\kappa_b = \frac{p_{o,b} - p_{e,b}}{1 - p_{e,b}},$$

where $p_{o,b}$ is the observed prequential accuracy in chunk $\mathcal{I}_b$, and $p_{e,b}$ is the expected agreement under the empirical label and prediction marginals in the same chunk. We report the chunk-averaged values $\overline{\mathrm{BA}}$ and $\overline{\kappa}$.

C.2 Fairness metrics

Fairness is evaluated using the profile-based protocol introduced in Section 4. At the final time T, we use the whole stream as the evaluation window and compute each fairness metric over the prefix sequence $\{\mathcal{R}^{\mathcal{F}}_\tau(T,T)\}_{\tau=1}^{T}$. Prefixes with fewer than 2000 prequential records are excluded from aggregation to avoid unstable early estimates. For a fairness metric ϕ, let $\mathcal{U}_\phi = \{\tau_1 < \cdots < \tau_m\}$ denote the remaining valid prefixes. The reported scalar fairness value is

$$\overline{\Phi}^{\mathcal{F}}_\phi = \sum_{i=1}^{m} \lambda_i \phi(\mathcal{R}^{\mathcal{F}}_{\tau_i}(T,T)), \qquad \lambda_i = \frac{\eta^{m-i}}{\sum_{j=1}^{m} \eta^{m-j}},$$

where we set $\eta = 0.999999$. This choice gives slightly larger weights to later prefixes while ensuring that the whole profile contributes to the final score.

All fairness metrics are reported as disparity scores, where smaller values indicate better fairness and 0 is optimal. For difference-based metrics, we use the absolute group gap

$$D(a,b) = |a - b|.$$

For ratio-based metrics, we use the min-symmetric deviation from parity,

$$R(a,b) = \left| 1 - \min\left\{ \frac{a}{b}, \frac{b}{a} \right\} \right|,$$

whenever the required denominators are valid. Ratio-based values are therefore also minimized at 0, corresponding to perfect parity.

Let $S \in \{0,1\}$ denote the sensitive attribute, $Y \in \{0,1\}$ the true label, and $\hat{Y} \in \{0,1\}$ the prequential prediction. For group g, we define

$$\mathrm{PPR}_g = P(\hat{Y} = 1 \mid S = g), \quad \mathrm{TPR}_g = P(\hat{Y} = 1 \mid Y = 1, S = g),$$
$$\mathrm{FNR}_g = P(\hat{Y} = 0 \mid Y = 1, S = g), \quad \mathrm{FPR}_g = P(\hat{Y} = 1 \mid Y = 0, S = g),$$
$$\mathrm{FOR}_g = P(Y = 1 \mid \hat{Y} = 0, S = g), \quad \mathrm{FDR}_g = P(Y = 0 \mid \hat{Y} = 1, S = g),$$
$$\mathrm{PPV}_g = P(Y = 1 \mid \hat{Y} = 1, S = g), \quad \mathrm{NPV}_g = P(Y = 0 \mid \hat{Y} = 0, S = g),$$

and

$$\mathrm{ERR}_g = P(\hat{Y} \neq Y \mid S = g).$$

All probabilities are empirical probabilities computed on the corresponding prequential prefix.

For individual-fairness and inequality metrics, let $b_i = \hat{y}_i - y_i + 1$ denote the benefit assigned to instance i, let $\mu = n^{-1} \sum_{i=1}^{n} b_i$, and let μ_g be the mean benefit in group g. We use the generalized entropy family with parameter α, and set $\alpha = 2$ unless otherwise specified. The Theil index is the limiting case $\alpha \to 1$, and the coefficient of variation is the corresponding $\alpha = 2$ dispersion form.

For compact notation, define $D(a,b) = |a - b|$ and $R(a,b) = |1 - \min\{a/b, b/a\}|$. Let $\mathrm{PPR}_g = P(\hat{Y} = 1 \mid S = g)$, $\mathrm{TPR}_g = P(\hat{Y} = 1 \mid Y = 1, S = g)$, $\mathrm{FNR}_g = P(\hat{Y} = 0 \mid Y = 1, S = g)$, $\mathrm{FPR}_g = P(\hat{Y} = 1 \mid Y = 0, S = g)$, $\mathrm{FOR}_g = P(Y = 1 \mid \hat{Y} = 0, S = g)$, $\mathrm{FDR}_g = P(Y = 0 \mid \hat{Y} = 1, S = g)$, $\mathrm{PPV}_g = P(Y = 1 \mid \hat{Y} = 1, S = g)$, $\mathrm{NPV}_g = P(Y = 0 \mid \hat{Y} = 0, S = g)$, and $\mathrm{ERR}_g = P(\hat{Y} \neq Y \mid S = g)$.

Table 3: Fairness metrics used in our evaluation. All metrics are computed on profile prefixes and aggregated with exponential decay. Smaller values indicate better fairness, and 0 is optimal.

Criterion	Abbr.	Metric name	Formulation
Independence	SPD	Statistical parity difference	$D(\mathrm{PPR}_0, \mathrm{PPR}_1)$
Independence	DI	Disparate impact	$R(\mathrm{PPR}_0, \mathrm{PPR}_1)$
Separation	FNR-D	False negative rate difference	$D(\mathrm{FNR}_0, \mathrm{FNR}_1)$
Separation	FNR-R	False negative rate ratio	$R(\mathrm{FNR}_0, \mathrm{FNR}_1)$
Separation	FPR-D	False positive rate difference	$D(\mathrm{FPR}_0, \mathrm{FPR}_1)$
Separation	FPR-R	False positive rate ratio	$R(\mathrm{FPR}_0, \mathrm{FPR}_1)$
Separation	AOD	Average odds difference	$\left\lvert \frac{1}{2}\left[(\mathrm{TPR}_0 - \mathrm{TPR}_1) + (\mathrm{FPR}_0 - \mathrm{FPR}_1)\right]\right\rvert$
Separation	AAOD	Average absolute odds difference	$\frac{1}{2}\left[D(\mathrm{TPR}_0, \mathrm{TPR}_1) + D(\mathrm{FPR}_0, \mathrm{FPR}_1)\right]$
Sufficiency	FOR-D	False omission rate difference	$D(\mathrm{FOR}_0, \mathrm{FOR}_1)$
Sufficiency	FOR-R	False omission rate ratio	$R(\mathrm{FOR}_0, \mathrm{FOR}_1)$
Sufficiency	FDR-D	False discovery rate difference	$D(\mathrm{FDR}_0, \mathrm{FDR}_1)$
Sufficiency	FDR-R	False discovery rate ratio	$R(\mathrm{FDR}_0, \mathrm{FDR}_1)$
Sufficiency	APD	Average predictive value difference	$\frac{1}{2}\left[D(\mathrm{PPV}_0, \mathrm{PPV}_1) + D(\mathrm{NPV}_0, \mathrm{NPV}_1)\right]$
Error parity	ER-D	Error rate difference	$D(\mathrm{ERR}_0, \mathrm{ERR}_1)$
Error parity	ER-R	Error rate ratio	$R(\mathrm{ERR}_0, \mathrm{ERR}_1)$
Bias amplification	DFBA	Differential fairness bias amplification	$\left\lvert \epsilon(\hat{Y}, S) - \epsilon(Y, S) \right\rvert$, where $\epsilon(Z, S) = \max_z \max_{g, g'} \lvert \log P(Z = z \mid S = g) - \log P(Z = z \mid S = g') \rvert$
Between-group inequality	BGEI	Between-group generalized entropy index	$\frac{1}{n\alpha(\alpha-1)} \sum_{g \in \{0,1\}} n_g \left[\left(\frac{\mu_g}{\mu} \right)^{\alpha} - 1 \right]$
Between-group inequality	BTI	Between-group Theil index	$\frac{1}{n} \sum_{g \in \{0,1\}} n_g \frac{\mu_g}{\mu} \log \frac{\mu_g}{\mu}$
Between-group inequality	BCV	Between-group coefficient of variation	$\sqrt{\frac{1}{n} \sum_{g \in \{0,1\}} n_g \left(\frac{\mu_g}{\mu} - 1 \right)^2}$
Individual fairness	GEI	Generalized entropy index	$\frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} \left[\left(\frac{b_i}{\mu} \right)^{\alpha} - 1 \right]$
Individual fairness	TI	Theil index	$\frac{1}{n} \sum_{i=1}^{n} \frac{b_i}{\mu} \log \frac{b_i}{\mu}$
Individual fairness	CV	Coefficient of variation	$\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{b_i}{\mu} - 1 \right)^2}$

D Datasets and Stream Construction

D.1 Overview

We construct synthetic data streams by separating three factors: the feature–label generation mechanism, the label–sensitive-attribute distribution, and the instance-level difficulty composition. Specifically, a base stream generator G first produces feature–label pairs (x_t, y_t). We then augment each instance with a binary sensitive attribute $s_t \in \{0, 1\}$ according to a controlled joint distribution over (Y, S). This construction preserves the feature-generation mechanism of the original stream while allowing us to vary class imbalance (CI), sensitive-group imbalance (SI), and data-side label–group dependence in a controlled manner. For IDI experiments, we additionally control the proportions of safe, borderline, rare, and outlier instances following the difficulty factors studied in imbalanced and drifting data streams (Brzezinski et al., 2021).

D.2 Base synthetic stream generators

We use three synthetic stream generators: Hyperplane (Hulten et al., 2001), RandomRBF (Bifet et al., 2010a), and Difficulty. All generated streams are stationary in the main experiments, so that the observed changes in fairness can be attributed to the controlled imbalance sources rather than concept drift.

HYPERPLANE generates instances in a d-dimensional feature space. Let $w = (w_1, \ldots, w_d)$ be a fixed weight vector with $w_i \sim U(0, 1)$. For each instance, features are sampled independently as $x_i \sim U(0, 1)$. The normalized hyperplane score is

$$r(x) = \frac{\sum_{i=1}^{d} w_i x_i}{\sum_{i=1}^{d} w_i}.$$

The class label is assigned by a fixed threshold,

$$y = \begin{cases} 1, & r(x) > 0.5, \\ 0, & r(x) \leq 0.5. \end{cases}$$

This generator provides a simple linear decision boundary and is used to study how CI and SI affect fairness under a relatively regular classification structure.

RANDOMRBF generates data from a mixture of radial basis components. Let K denote the number of centroids. The j-th centroid is defined by

$$C_j = (c_j, \sigma_j, y_j, \alpha_j),$$

where $c_j \in \mathbb{R}^d$ is the centroid location, $\sigma_j > 0$ is the standard deviation, $y_j \in \{0, 1\}$ is the class label associated with the centroid, and α_j is its sampling weight. To generate an instance, a centroid C_j is sampled according to α_j. A random unit direction $\hat{u}$ is then drawn, a displacement length $r \sim N(0, \sigma_j)$ is sampled, and the feature vector is generated as

$$x = c_j + r\hat{u}.$$

The output label is y_j. This generator creates nonlinear class regions and is used to test whether the effects of CI and SI persist beyond linear decision boundaries.

DIFFICULTY is used to control instance-level difficulty imbalance (IDI). Following the safe/borderline/rare/outlier taxonomy for difficult examples (Napierala & Stefanowski, 2016; Brzezinski et al., 2021), each stream is constructed from a mixture of four instance types:

$$\pi_{\text{safe}} + \pi_{\text{border}} + \pi_{\text{rare}} + \pi_{\text{out}} = 1.$$

Safe instances are sampled from locally class-consistent regions. Borderline instances are sampled near class boundaries, where local neighborhoods contain mixed labels. Rare instances form small class-consistent subregions with limited support, while outlier instances are isolated or locally inconsistent samples. By changing the mixture vector

$$\boldsymbol{\pi}_{\text{IDI}} = (\pi_{\text{safe}}, \pi_{\text{border}}, \pi_{\text{rare}}, \pi_{\text{out}}),$$

we obtain streams with different IDI profiles while keeping the remaining stream properties fixed.

D.3 CONTROLLING CI, SI, AND DATA-SIDE FAIRNESS

Let

$$q = P(Y = 1), \qquad r = P(S = 1),$$

where q controls the class marginal and r controls the sensitive-group marginal. For binary labels and binary sensitive attributes, CI and SI are measured by

$$\rho_{\text{CI}} = \min\{q, 1 - q\}, \qquad \rho_{\text{SI}} = \min\{r, 1 - r\}.$$

Smaller values indicate stronger imbalance.

To control data-side label–group dependence, we specify a target label-side statistical parity gap

$$\Delta_{\text{data}}^{\text{SPD}} = |P(Y = 1 \mid S = 0) - P(Y = 1 \mid S = 1)|.$$

Given q, r, and a target gap Δ, we construct a feasible joint distribution over (Y, S) as follows. Let $\zeta \in \{-1, +1\}$ determine which sensitive group has the larger positive-label rate, and define

$$a_1 = P(Y = 1 \mid S = 1) = q + (1 - r)\zeta\Delta,$$
$$a_0 = P(Y = 1 \mid S = 0) = q - r\zeta\Delta.$$

The resulting joint probabilities are

$$P(Y=1,S=1)=ra_1, \qquad P(Y=0,S=1)=r(1-a_1),$$
$$P(Y=1,S=0)=(1-r)a_0, \qquad P(Y=0,S=0)=(1-r)(1-a_0).$$

We keep only feasible settings for which $a_0, a_1 \in [0,1]$. This construction preserves the desired marginals $P(Y=1)=q$ and $P(S=1)=r$, while setting

$$|a_0 - a_1| = \Delta.$$

Thus, CI, SI, and data-side fairness can be varied without confounding their target definitions.

Given a generated feature–label pair (x_t, y_t), we assign the sensitive attribute by

$$S_t \mid Y_t = 1 \sim \text{Bernoulli}(\theta_1), \qquad S_t \mid Y_t = 0 \sim \text{Bernoulli}(\theta_0),$$

where

$$\theta_1 = P(S=1 \mid Y=1) = \frac{P(Y=1,S=1)}{P(Y=1)},$$

and

$$\theta_0 = P(S=1 \mid Y=0) = \frac{P(Y=0,S=1)}{P(Y=0)}.$$

When $\Delta = 0$, we have $P(Y=1 \mid S=0) = P(Y=1 \mid S=1)$, and therefore Y and S are independent. This corresponds to data-side fair streams. When $\Delta > 0$, the sensitive attribute is statistically associated with the label, producing data-side unfair streams.

The commonly used Direct construction is a special case of this formulation. If $q = 0.5$, $r = 0.5$, and

$$P(S=1 \mid Y=1) = p, \qquad P(S=1 \mid Y=0) = 1-p,$$

then

$$\Delta^{\text{SPD}}_{\text{data}} = |2p-1|.$$

Hence $p = 0.5$ yields $Y \perp S$, while $p \neq 0.5$ introduces controlled label–group dependence.

D.4 Diagnostic imbalance grids

For the motivating analysis and RQ1, we use diagnostic grids that vary one imbalance source at a time while respecting the feasibility constraint of the joint distribution over (Y, S). Let

$$\alpha = P(Y=1), \qquad \beta = P(S=1),$$

and define the data-side label–group dependence parameter

$$d = P(Y=1 \mid S=1) - P(Y=1 \mid S=0).$$

Given (α, β, d), we parameterize the conditional label rates by

$$P(Y=1 \mid S=1) = \alpha + (1-\beta)d, \qquad P(Y=1 \mid S=0) = \alpha - \beta d.$$

This construction preserves the marginal class probability $P(Y=1)=\alpha$. It is feasible only when both conditional probabilities lie in $[0,1]$, i.e.,

$$\max\left\{\frac{\alpha-1}{\beta}, -\frac{\alpha}{1-\beta}\right\} < d < \min\left\{\frac{\alpha}{\beta}, \frac{1-\alpha}{1-\beta}\right\}.$$

The label-side statistical parity gap is therefore

$$\Delta^{\text{SPD}}_{\text{data}} = |d|.$$

For data-side fair streams, we set $d = 0$. To study SI, we keep the class distribution balanced with $\alpha = 0.5$ and vary

$$\beta \in \{0.50, 0.10, 0.05, 0.02, 0.01\}.$$

To study CI, we keep the sensitive attribute balanced with $\beta = 0.5$ and vary

$$\alpha \in \{0.50, 0.10, 0.05, 0.02, 0.01\}.$$

Since $d = 0$, all these settings are feasible.

For data-side unfair streams, we consider

$$|d| \in \{0.05, 0.10, 0.15, 0.20, 0.25\}.$$

When studying SI under balanced labels, we fix $\alpha = 0.5$ and vary

$$\beta \in \{0.50, 0.10, 0.05, 0.02, 0.01\}.$$

These settings are feasible for the above d-values because the constraint reduces to

$$|d| < \frac{1}{2\max\{\beta, 1-\beta\}}.$$

When studying CI under balanced sensitive groups, we fix $\beta = 0.5$. In this case the feasibility constraint becomes

$$|d| < 2\min\{\alpha, 1-\alpha\}.$$

Therefore, to keep the same unfairness grid up to $|d| = 0.25$, we use the feasible class-imbalance grid

$$\alpha \in \{0.50, 0.45, 0.35, 0.25, 0.15\},$$

rather than the more extreme fair-stream grid $\{0.50, 0.10, 0.05, 0.02, 0.01\}$. This avoids invalid joint distributions while still covering increasingly imbalanced class marginals.

Given the sampled label y_t, the sensitive attribute is generated from the corresponding conditional distribution implied by the joint table:

$$P(S = 1 \mid Y = 1) = \frac{\beta[\alpha + (1-\beta)d]}{\alpha},$$

and

$$P(S = 1 \mid Y = 0) = \frac{\beta[1 - \alpha - (1-\beta)d]}{1-\alpha}.$$

Thus, the generated stream preserves the target α, β, and d in expectation.

D.5 IDI DIAGNOSTIC GRIDS

In addition to CI and SI, we construct diagnostic streams for instance-level difficulty imbalance (IDI) using DIFFICULTY. Following the safe/borderline/rare/outlier taxonomy of difficult examples (Brzezinski et al., 2021), each stream is specified by the mixture

$$\boldsymbol{\pi}_{\mathrm{IDI}} = (\pi_{\mathrm{safe}}, \pi_{\mathrm{border}}, \pi_{\mathrm{rare}}, \pi_{\mathrm{out}}), \qquad \pi_{\mathrm{safe}} + \pi_{\mathrm{border}} + \pi_{\mathrm{rare}} + \pi_{\mathrm{out}} = 1.$$

We vary the proportion of the dominant difficult instance type by

$$\rho_{\mathrm{IDI}} \in \{0.0, 0.2, 0.4, 0.6, 0.8\}.$$

For borderline-dominated streams, we set

$$\pi_{\mathrm{border}} = \rho_{\mathrm{IDI}}, \qquad \pi_{\mathrm{safe}} = 1 - \rho_{\mathrm{IDI}}, \qquad \pi_{\mathrm{rare}} = \pi_{\mathrm{out}} = 0.$$

Rare- and outlier-dominated streams are constructed analogously by assigning ρ_{IDI} to π_{rare} or π_{out}, respectively. Thus, $\rho_{\mathrm{IDI}} = 0$ corresponds to a safe-only stream, while larger values indicate stronger dominance of a specific difficult instance type.

This design separates IDI severity from IDI type. The parameter ρ_{IDI} controls how many high-difficulty instances appear in the stream, whereas the selected dominant component controls the nature of the difficulty. Borderline instances mainly introduce local class ambiguity near decision boundaries, rare instances provide sparse but locally coherent support, and outlier instances introduce isolated or locally inconsistent evidence. This allows us to test whether fairness degradation is driven only by the amount of difficulty or also by the type of difficult examples dominating the stream.

D.6 REPRESENTATIVE MULTI-SOURCE IMBALANCE SETTINGS

Beyond the diagnostic grids, we construct 19 representative settings for comprehensive evaluation. These settings cover balanced streams, single-source imbalance, pairwise imbalance, and joint CI–SI–IDI imbalance. CI and SI are controlled by the minority class and minority sensitive-group proportions, respectively, while IDI is controlled by the mixture of safe, borderline, rare, and outlier instances. The resulting settings allow us to evaluate whether fairness-aware streaming methods remain reliable when multiple imbalance sources appear simultaneously, rather than only under isolated CI, SI, or IDI. The full list is reported in Table 4.

Table 4: Representative experimental settings for multi-source imbalance. CI and SI are controlled by ρ_{CI} and ρ_{SI}, respectively; IDI is controlled by the proportions of safe (s), borderline (b), rare (p), and outlier (o) instances. Symbols indicate imbalance strength: ○ denotes balanced/absent, ◐ denotes strong imbalance, and ● denotes severe imbalance. For CI/SI, ○ $= 0.50$, ◐ $= 0.35$, and ● $= 0.15$; for B/R/O, ○ means the component is absent, ◐ means it is active in a strong IDI setting, and ● means it is active in the full stress setting.

Idx.	#Src.	CI	SI	B	R	O	ρ_{CI}	ρ_{SI}	s	b	p	o	Setting
S0	0	○	○	○	○	○	0.50	0.50	1.00	0	0	0	Balanced baseline
S1	1	◐	○	○	○	○	0.35	0.50	1.00	0	0	0	Strong CI only
S2	1	●	○	○	○	○	0.15	0.50	1.00	0	0	0	Severe CI only
S3	1	○	◐	○	○	○	0.50	0.35	1.00	0	0	0	Strong SI only
S4	1	○	●	○	○	○	0.50	0.15	1.00	0	0	0	Severe SI only
S5	1	○	○	◐	○	○	0.50	0.50	0.40	0.60	0	0	IDI only: borderline
S6	1	○	○	○	◐	○	0.50	0.50	0.40	0	0.60	0	IDI only: rare
S7	1	○	○	○	○	◐	0.50	0.50	0.40	0	0	0.60	IDI only: outlier
S8	1	○	○	◐	◐	◐	0.50	0.50	0.40	0.20	0.20	0.20	IDI only: mixed
S9	2	◐	◐	○	○	○	0.35	0.35	1.00	0	0	0	Strong CI + SI
S10	2	●	●	○	○	○	0.15	0.15	1.00	0	0	0	Severe CI + SI
S11	2	◐	○	◐	◐	◐	0.35	0.50	0.40	0.20	0.20	0.20	CI + mixed IDI
S12	2	○	◐	◐	◐	◐	0.50	0.35	0.40	0.20	0.20	0.20	SI + mixed IDI
S13	3	◐	◐	◐	○	○	0.35	0.35	0.40	0.60	0	0	Joint imbalance + borderline
S14	3	◐	◐	○	◐	○	0.35	0.35	0.40	0	0.60	0	Joint imbalance + rare
S15	3	◐	◐	○	○	◐	0.35	0.35	0.40	0	0	0.60	Joint imbalance + outlier
S16	3	◐	◐	◐	◐	◐	0.35	0.35	0.40	0.20	0.20	0.20	Strong CI + SI + mixed IDI
S17	3	●	●	◐	◐	◐	0.15	0.15	0.40	0.20	0.20	0.20	Severe CI/SI + mixed IDI
S18	3	●	●	●	●	●	0.15	0.15	0.20	0.27	0.27	0.26	Full stress setting

D.7 Real-world streams

In addition to synthetic streams, we evaluate on five real-world datasets: Adult (Becker & Kohavi, 1996), COMPAS (Larson et al., 2016), KDD (kdd, 2000), Census (Kohavi, 1996), and Dutch (Kamiran & Calders, 2012). Each dataset is converted into a data stream by ordering or randomly permuting the preprocessed instances and then applying the same prequential evaluation protocol as for synthetic streams. No artificial resampling is applied when constructing the real-world streams; their observed class imbalance, sensitive-group imbalance, and data-side label–group dependence are preserved. These datasets therefore serve as external-validity checks for the controlled synthetic findings.

Categorical features are encoded consistently across methods, continuous features are normalized when required by the corresponding learner, and the sensitive attribute is removed from the feature vector only for methods whose original protocol requires this restriction. Unless otherwise specified, all methods receive the same stream order and the same preprocessed feature representation for a given dataset.

D.8 Synthetic-to-real Profile and Behavior Consistency

We pair each real-world stream with a synthetic setting having a similar CI/SI/IDI profile among S0–S18. CI and SI are represented by the minority class and sensitive-group proportions, respectively, while the IDI profile records the positive-class proportions of safe, borderline, rare, and outlier instances. These profiles are used for descriptive matching rather than to assume that a real stream follows the synthetic generator. Table 5 then compares the behavior of all seven methods within each pair. For each metric in the order BA/Kappa/SPD/DI/GM_4, $x/6$ counts the baselines against which WELL has the same win/loss outcome in the real and synthetic streams. "WELL–base" aggregates these outcomes over five metrics, giving 30 comparisons, while "All-method" counts consistent orderings over all 21 method pairs and five metrics, giving 105 comparisons. WELL–base agreement ranges from 77% to 87%, and All-method agreement ranges from 69% to 88%; four of the five pairs reach at least 80% All-method agreement, while Census–S11 is lower at 69%. The agreement is therefore broad but not universal, providing direct evidence that the source-specific patterns observed under controlled CI/SI/IDI settings are also informative in real-world streams.

Table 5: Synthetic-to-real profile matching and performance agreement. Per-metric entries follow BA/Kappa/SPD/DI/GM_4.

Real–synthetic pair	CI	SI	IDI (S/B/R/O)	Per metric	WELL–base	All-method
Adult–S16	0.25	0.33	0.35 / 0.39 / 0.15 / 0.10	4/6, 6/6, 5/6, 5/6, 6/6	26/30 (87%)	92/105 (88%)
COMPAS–S12	0.47	0.19	0.34 / 0.45 / 0.15 / 0.06	5/6, 3/6, 6/6, 6/6, 6/6	26/30 (87%)	90/105 (86%)
KDD–S18	0.06	0.48	0.06 / 0.34 / 0.29 / 0.31	6/6, 3/6, 5/6, 5/6, 5/6	24/30 (80%)	84/105 (80%)
Census–S11	0.06	0.39	0.23 / 0.36 / 0.21 / 0.19	5/6, 4/6, 4/6, 5/6, 5/6	23/30 (77%)	72/105 (69%)
Dutch–S5	0.48	0.49	0.62 / 0.27 / 0.07 / 0.04	6/6, 5/6, 5/6, 5/6, 5/6	26/30 (87%)	88/105 (84%)

E Experimental Protocol and Compared Methods

E.1 Compute Resources

The main accuracy–fairness experiments were conducted on CPU using a server equipped with an AMD EPYC 9654 96-Core Processor. GPU was not used in these experiments. Each individual run took approximately 10 minutes on a single CPU core, and each experimental setting was independently repeated ten times. The separate runtime benchmark and its hardware are described in Appendix I.1.1. No additional preliminary experiments were included in the reported compute budget.

E.2 Shared Prequential Evaluation Protocol

All experiments follow a unified prequential test-then-train protocol. Given a stream $\mathcal{D} = \{(x_t, s_t, y_t)\}_{t=1}^{T}$, each method first predicts the label of the incoming instance before observing its ground-truth label:

$$\hat{y}_t = f_t(x_t, s_t).$$

After the prediction is recorded, the true label y_t is revealed and the method updates its model using the current instance. Thus, every reported prediction is made by a model that has not yet been trained on the same instance, which avoids evaluation leakage and matches the standard online stream-learning setting.

All accuracy and fairness metrics are computed from this record. Accuracy metrics are computed over consecutive stream chunks following (Aguiar et al., 2024), while fairness metrics are computed under the profile-based protocol described in Appendix C. Unless otherwise specified, figures and tables report the mean value over the 10 times. When rank-based summaries are used, ranks are first computed within each stream setting and then averaged over settings, so that no single stream configuration dominates the comparison.

All algorithms process each stream in a single pass, and no method is allowed to use future instances, future labels, or test-time statistics unavailable at the current time step. Hyperparameters of compared methods follow their original papers or official implementations. When an original method provides dataset-independent default parameters, we use those defaults without tuning on our evaluation streams. When multiple parameter choices are reported in the original work, we use the recommended streaming configuration. This protocol is intended to compare the behavior of different methods under the same CI, SI, and IDI conditions, rather than to optimize each method separately for each imbalance setting.

All methods use the same preprocessed feature representation for a given dataset. The sensitive attribute is retained or removed from the model input according to the original protocol of each compared method, but it is always kept for evaluation so that group-based fairness metrics can be computed from the same prequential prediction records. For WELL and its ablations, the same protocol is used, with only the specified weighting component removed in each ablated variant.

E.3 Statistical Analysis for the Motivating Analysis

Cross-stream trend estimation. The motivating analysis studies how each imbalance source changes algorithmic behavior across controlled stationary streams. For a fixed generator, data-side fairness setting, algorithm, and metric, we collect the metric values from a sequence of streams in which one imbalance source is varied while the remaining factors are fixed. Let $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(R)}$

denote these streams ordered from weakest to strongest imbalance, and let m_r be the corresponding metric value on $\mathcal{D}^{(r)}$. We assign each stream an ordinal imbalance index

$$v_r = r - 1, \qquad r = 1, \ldots, R,$$

so that larger v_r consistently indicates stronger imbalance.

For each algorithm–metric pair, we fit a simple linear trend model

$$m_r = a v_r + b + \varepsilon_r,$$

where a is the estimated cross-stream slope. The slope a is used as the cell value in the trend heatmaps. For fairness disparity metrics, such as SPD, DI, and other group-gap metrics, larger values indicate worse fairness; therefore, a positive slope means that fairness degrades as imbalance becomes stronger. For accuracy metrics, where larger values indicate better performance, we use the corresponding error form, such as $1 - \mathrm{BA}$ and $1 - \kappa$, so that positive slopes consistently indicate performance degradation. Thus, the magnitude of a reflects how strongly a metric changes across imbalance levels, while its sign indicates whether the metric improves or worsens under stronger imbalance.

Metric-correlation analysis We further analyze whether different metrics give consistent conclusions across imbalanced streams. For each fixed scenario, we construct metric-pair point sets from all algorithms and all repeated runs. Given two metrics M_i and M_j, let

$$\mathcal{P}_{ij} = \left\{ \left(M_i^{(q)}, M_j^{(q)} \right) \right\}_{q=1}^{Q}$$

be the collection of paired metric values, where q indexes an algorithm–run result under the considered stream setting. Before comparing metrics, all quantities are oriented so that smaller values are better; for example, BA and Kappa are converted to $1 - \mathrm{BA}$ and $1 - \kappa$, while fairness gaps are kept unchanged.

To focus the correlation analysis on practically relevant trade-off regions, we first extract the non-dominated subset

$$\mathcal{P}_{ij}^{\star} = \{ p \in \mathcal{P}_{ij} : \nexists p' \in \mathcal{P}_{ij} \text{ such that } p'_1 \le p_1,\ p'_2 \le p_2, \text{ and } p' \ne p \}.$$

This set corresponds to the empirical Pareto frontier for the two metrics. We then divide the metric plane into a 5×5 grid using the value ranges of $\mathcal{P}_{ij}$. Let $G(p)$ denote the grid cell containing point p. We collect all grid cells that contain at least one non-dominated point,

$$\mathcal{G}_{ij}^{\star} = \left\{ G(p) : p \in \mathcal{P}_{ij}^{\star} \right\},$$

and define the grid-expanded trade-off set as

$$\widetilde{\mathcal{P}}_{ij} = \left\{ p \in \mathcal{P}_{ij} : G(p) \in \mathcal{G}_{ij}^{\star} \right\}.$$

This procedure keeps the local neighborhoods around the empirical Pareto frontier rather than using only the frontier points themselves, which reduces the instability caused by a small number of non-dominated samples.

Finally, we compute the Spearman rank correlation between M_i and M_j on $\widetilde{\mathcal{P}}_{ij}$:

$$\rho_{ij}^{\mathrm{S}} = \mathrm{corr}\left(\mathrm{rank}\big(\{M_i(p) : p \in \widetilde{\mathcal{P}}_{ij}\}\big), \mathrm{rank}\big(\{M_j(p) : p \in \widetilde{\mathcal{P}}_{ij}\}\big) \right).$$

The resulting matrix ρ_{ij}^{S} is visualized as the metric-correlation heatmap. A positive value indicates that two metrics tend to rank algorithms similarly in the relevant trade-off region, whereas a negative value indicates that improving one metric may conflict with the other.

E.4 Algorithms Used in the Motivating Analysis

The motivating analysis evaluates 15 online data-stream classifiers under the controlled CI, SI, and IDI stream settings described in Appendix D. The compared algorithms are GAST (Assis et al., 2024), SGBT (Gunasekara et al., 2024), ARFR (Ferreira et al., 2019), IRL (Bernardo et al., 2020), SMOTEOB (Bernardo & Della Valle, 2021a), VFC-SMOTE (Bernardo & Della Valle, 2021b), CSARF (Loezer et al., 2020), GHVFDT (LYON et al., 2014), KUE (Cano & Krawczyk, 2020),

Table 6: WELL ablations used in RQ2. Each variant removes one component of the instance-weighting rule while keeping the remaining protocol unchanged.

Ablation	Removed component	Implementation
WELL w/o ℓ_t^{imb}	Group-imbalance weighting	$\beta = 0$
WELL w/o ℓ_t^{dist}	Distributional-association term	$\gamma_{\mathrm{dist}} = 0$
WELL w/o ℓ_t^{dec}	Decision-disparity term	$\gamma_{\mathrm{dec}} = 0$
WELL w/o ℓ_t^{fair}	All fairness-related terms	$\gamma_{\mathrm{dist}} = \gamma_{\mathrm{dec}} = 0$
WELL w/o d_t	Instance-level difficulty moderation	$\chi(d_t) = 1$

LB (Bifet et al., 2010b), ROSE (Cano & Krawczyk, 2022), ARF (Gomes et al., 2017), CALMID (Liu et al., 2021), UOB (Wang et al., 2016) and OOB (Wang et al., 2016). Among them, GAST, SGBT, and ARF are used as representative imbalance-agnostic stream learners, while the remaining methods are class-imbalance-aware stream classifiers selected from the comparisons summarized by (Aguiar et al., 2024).

All algorithms are implemented in MOA and evaluated using the test-then-train prequential protocol in Appendix E. We use the parameter settings recommended in the original papers or official MOA implementations. Accuracy and fairness are computed using the metrics described in Appendix C.

E.5 Fairness-aware Methods Used in RQ1

RQ1 evaluates six fairness-aware data-stream methods under the controlled CI, SI, and IDI stream settings described in Appendix D. The compared algorithms are CFDS (Ghosh et al., 2026), CFSMOTE (Lammers et al., 2025), FOMOS (Pham et al., 2024), FairPCA (Lee et al., 2023), FABBOO (Iosifidis & Ntoutsi, 2020), and FAHT (Zhang & Ntoutsi, 2019). These methods cover data-level, algorithm-level, and ensemble-based fairness methods. Among them, CFDS, FOMOS, FairPCA, and FAHT are not primarily designed to model CI, SI, and IDI as joint imbalance sources. FABBOO and CFSMOTE include imbalance-oriented mechanisms, but they do not explicitly address instance-level difficulty imbalance.

E.6 Methods and Ablations Used in RQ2

RQ2 compares WELL with the same six fairness-aware data-stream methods used in RQ1, namely CFDS (Ghosh et al., 2026), CFSMOTE (Lammers et al., 2025), FOMOS (Pham et al., 2024), FairPCA (Lee et al., 2023), FABBOO (Iosifidis & Ntoutsi, 2020), and FAHT (Zhang & Ntoutsi, 2019). These baselines provide a consistent reference for evaluating whether WELL improves the accuracy–fairness trade-off beyond existing fairness-aware streaming methods under CI, SI, and IDI.

The full WELL method uses the instance weight

$$\omega_t = \exp\big(\beta \ell_t^{\mathrm{imb}} + \gamma_{\mathrm{dist}} \ell_t^{\mathrm{dist}} + \gamma_{\mathrm{dec}} a_t^{\mathrm{dec}} \ell_t^{\mathrm{dec}}\big) \cdot \chi(d_t) \tag{56}$$

We further evaluate five ablated variants to isolate the contribution of each weighting component. Each ablation changes only the corresponding coefficient in Eq. equation 56, while keeping all other settings unchanged. The variants are summarized in Table 6.

Here ℓ_t^{fair} is a shorthand for the two fairness-related log-weight terms, i.e., $\{\ell_t^{\mathrm{dist}}, \ell_t^{\mathrm{dec}}\}$. Therefore, WELL w/o ℓ_t^{fair} keeps only the group-imbalance term and removes both distributional-association term and decision-disparity term. WELL w/o d_t keeps all imbalance and fairness terms but removes the difficulty-moderation penalty. This set of ablations tests whether WELL's performance comes from a single dominant term or from the joint effect of group-imbalance weighting, distributional correction, decision-disparity correction, and instance-level difficulty moderation.

E.7 Ranking and GM_4 Computation

For compact comparison, we report both metric-wise ranks and an overall accuracy–fairness trade-off score, denoted by GM_4. Ranks are computed separately within each stream setting. For BA and Kappa, larger values are better; for SPD, DI, and GM_4, smaller values are better. After ranking methods within each setting, we average the ranks across settings. This prevents settings with larger

metric magnitudes from dominating the summary. What's more, following (Zhang et al., 2023), GM_4 is defined as the fourth-root geometric mean of two accuracy-error terms, $1-\mathrm{BA}$ and $1-\mathrm{Kappa}$, and two fairness-gap terms, so that all components are minimization-oriented.

F Additional Results for the Motivating Analysis

F.1 Overview

This section supplements the motivating analysis in Section 5. The main text reports the core effects of CI, SI, and IDI on algorithmic fairness, while this appendix provides more detailed supporting evidence and additional diagnostics. We first analyze how imbalance reshapes metric agreement, which explains the metric choices used in later experiments. We further report two additional findings. All analyses use the shared prequential protocol and metrics described in Appendix E and Appendix C.

F.2 Preliminaries

We use SPD to examine group-conditional prediction disparity and distinguish it from differences in prediction counts. For a valid prequential prefix $\mathcal{R}_\tau^{\mathcal{F}}(t,w)$, let $\widehat{P}_\tau(\cdot)$ denote the empirical probability computed on this prefix. The SPD value at prefix τ is

$$\phi_{\mathrm{SPD}}\big(\mathcal{R}_\tau^{\mathcal{F}}(t,w)\big) = \Big|\widehat{P}_\tau(\hat{Y}=1 \mid S=1) - \widehat{P}_\tau(\hat{Y}=1 \mid S=0)\Big|. \tag{57}$$

For each sensitive group $S=s$, we decompose the positive prediction rate as

$$\widehat{P}_\tau(\hat{Y}=1 \mid S=s) = \widehat{P}_\tau(\hat{Y}=1, Y=0 \mid S=s) + \widehat{P}_\tau(\hat{Y}=1, Y=1 \mid S=s). \tag{58}$$

Thus,

$$\phi_{\mathrm{SPD}}\big(\mathcal{R}_\tau^{\mathcal{F}}(t,w)\big) = \big|\Delta_{\mathrm{FP}}^{\mathcal{F}}(\tau;t,w) + \Delta_{\mathrm{TP}}^{\mathcal{F}}(\tau;t,w)\big|, \tag{59}$$

where

$$\Delta_{\mathrm{FP}}^{\mathcal{F}}(\tau;t,w) = \widehat{P}_\tau(\hat{Y}=1, Y=0 \mid S=1) - \widehat{P}_\tau(\hat{Y}=1, Y=0 \mid S=0), \tag{60}$$

$$\Delta_{\mathrm{TP}}^{\mathcal{F}}(\tau;t,w) = \widehat{P}_\tau(\hat{Y}=1, Y=1 \mid S=1) - \widehat{P}_\tau(\hat{Y}=1, Y=1 \mid S=0). \tag{61}$$

These group-normalized components, abbreviated as $\Delta(\mathrm{FP})$ and $\Delta(\mathrm{TP})$, use each group's own sample count as the denominator. They are contributions to the positive prediction rate, not the standard FPR or TPR. Equation (59) holds for any relative group sizes. The common-reference count contrasts in the SI diagnostics below are distinct quantities.

Each value of $\phi_{\mathrm{SPD}}(\mathcal{R}_\tau^{\mathcal{F}}(t,w))$ is one coordinate of the profile $\mathbf{P}_{\phi_{\mathrm{SPD}}}^{\mathcal{F}}(t,w)$. The reported scalar SPD is obtained by aggregating these profile values as described in Appendix B.1.

Common-reference count contrasts. The middle panels of Figures 16–18 retain group representation through a common normalization. For the same valid prefix, let $N_{\tau,s} > 0$ be the number of observations in group s, and let $TP_{\tau,s}$ and $FP_{\tau,s}$ be its true-positive and false-positive counts. With $S=0$ as the reference group, define

$$C_{\mathrm{TP}}(\tau) = \frac{TP_{\tau,1} - TP_{\tau,0}}{N_{\tau,0}}, \qquad C_{\mathrm{FP}}(\tau) = \frac{FP_{\tau,1} - FP_{\tau,0}}{N_{\tau,0}}. \tag{62}$$

Writing $p_{\tau,s} = \widehat{P}_\tau(\hat{Y}=1 \mid S=s)$, their sum separates a scaled conditional-rate difference from a group-size contribution:

$$\begin{aligned} C_{\mathrm{TP}}(\tau) + C_{\mathrm{FP}}(\tau) &= \frac{N_{\tau,1}}{N_{\tau,0}} p_{\tau,1} - p_{\tau,0} \\ &= \frac{N_{\tau,1}}{N_{\tau,0}}(p_{\tau,1} - p_{\tau,0}) + \left(\frac{N_{\tau,1}}{N_{\tau,0}} - 1\right) p_{\tau,0}. \end{aligned} \tag{63}$$

Thus, fewer positive predictions from a smaller group can produce a large count contrast even when the two conditional rates are similar. When $N_{\tau,1} = N_{\tau,0}$, the absolute sum reduces to SPD. Under SI, the count contrasts describe prediction volume, while the group-conditional rates and SPD assess prediction disparity.

F.3 Detailed evidence for Finding 1: CI exposes algorithmic unfairness through group-wise error disparities.

We further diagnose the CI results using fair streams. We vary only the class-imbalance degree $\rho_{\mathrm{CI}} \in \{0.50, 0.05, 0.01\}$, keep the sensitive attribute balanced, and set $y = 1$ as the minority class when $\rho_{\mathrm{CI}} < 0.50$. For each setting, we evaluate the same 15 online stream classifiers used in the motivating analysis.

Figures 9 and 10 first show the overall behavior of SPD and DI under CI. In terms of SPD, most algorithms remain close to parity under fair stream data, but their curves still change across CI settings. This suggests that CI may not always create large absolute SPD gaps, but it can still affect the group-wise prediction patterns used to compute SPD. In contrast, DI is more sensitive to severe CI. This is because SPD measures the absolute gap between $P(\hat{Y} = 1 \mid S = 0)$ and $P(\hat{Y} = 1 \mid S = 1)$, whereas DI depends on their ratio. When positive prediction rates become small under stronger CI, a small absolute gap can still correspond to a larger relative gap.

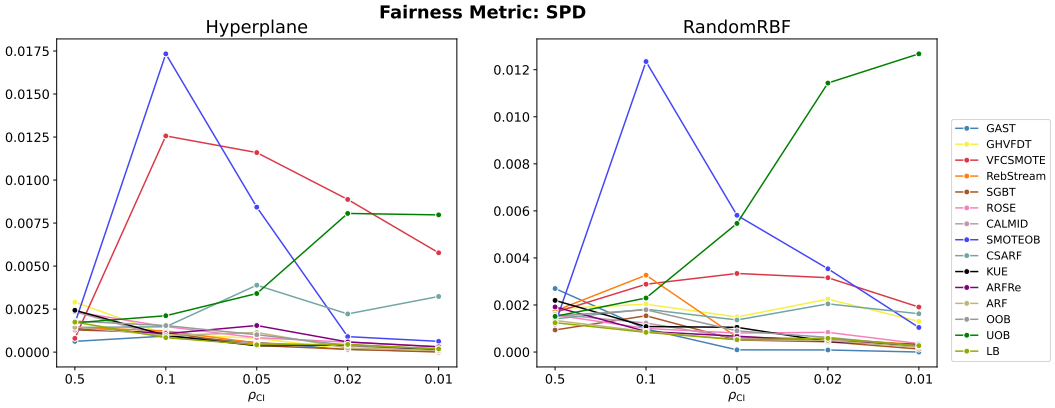


Figure 9: SPD of 15 online classifiers under fair Hyperplane and RandomRBF streams with different CI degrees.

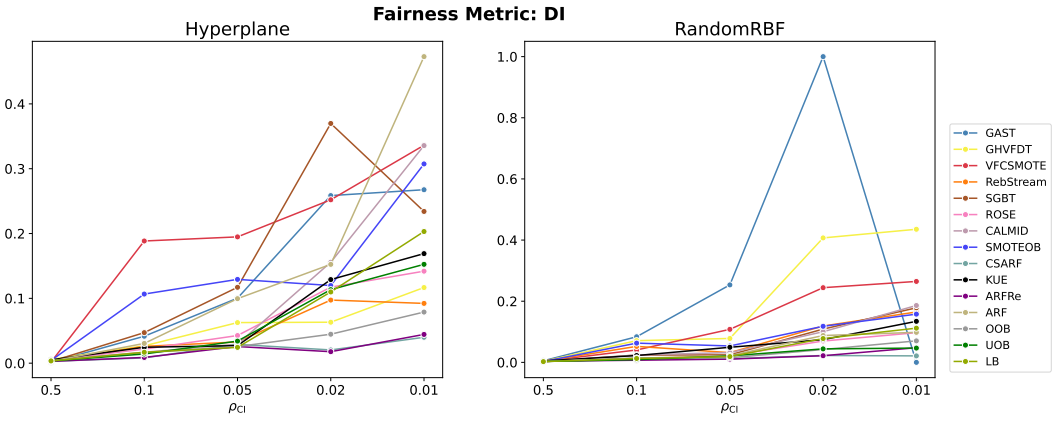


Figure 10: DI of 15 online classifiers under fair Hyperplane and RandomRBF streams with different CI degrees.

Figures 11–13 further decompose SPD on fair Hyperplane streams. Each row corresponds to one classifier. The first column shows $P(\hat{Y} = 1 \mid S = 0)$ and $P(\hat{Y} = 1 \mid S = 1)$, the second column shows $\Delta(\mathrm{TP})$ and $\Delta(\mathrm{FP})$, and the third column shows the resulting SPD profile. Under stronger CI, the positive-label mass becomes smaller for both sensitive groups. Accordingly, the terms $P(\hat{Y} = 1, Y = 1 \mid S = s)$ and $P(\hat{Y} = 1, Y = 0 \mid S = s)$ also shrink in scale. In the

decomposition plots, $\Delta(\mathrm{TP}) = \Delta P(\hat{Y} = 1, Y = 1 \mid S)$ is often close to zero, while the remaining SPD variation is more visibly associated with $\Delta(\mathrm{FP}) = \Delta P(\hat{Y} = 1, Y = 0 \mid S)$. Since the SPD profile is

$$|\Delta(\mathrm{TP}) + \Delta(\mathrm{FP})|\,,$$

the reduction of the overall positive prediction level helps explain why SPD can remain small under CI. At the same time, the same reduction makes DI more sensitive to small group-wise differences in positive prediction rates.

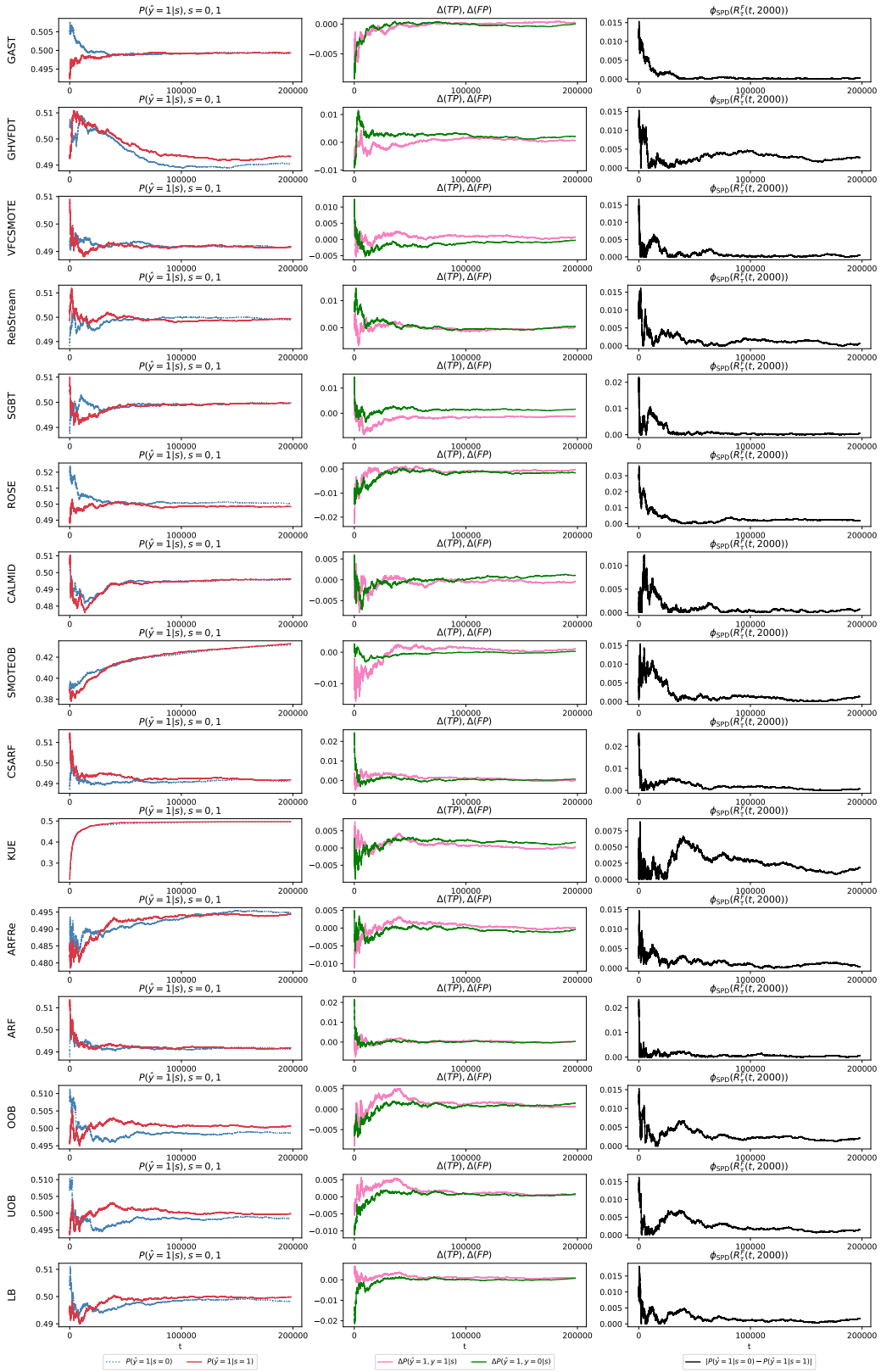


Figure 11: SPD decomposition on fair HYPERPLANE streams with $\rho_{\mathrm{CI}} = 0.50$ over 15 online classifiers. The first column shows $P(\hat{Y}=1 \mid S=0)$ and $P(\hat{Y}=1 \mid S=1)$; the second column shows $\Delta(\mathrm{TP})$ and $\Delta(\mathrm{FP})$, whose absolute sum gives the SPD profile.

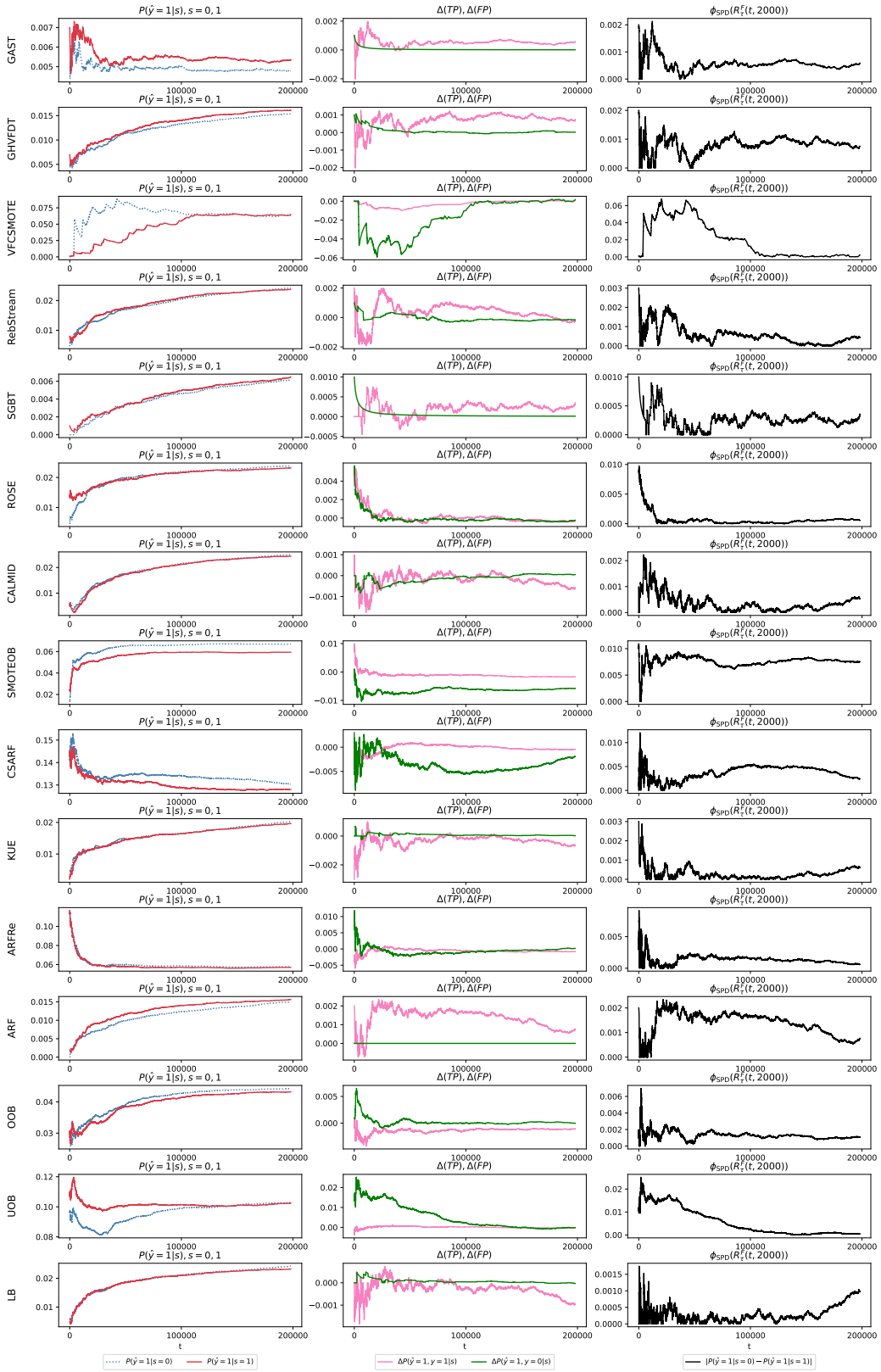


Figure 12: SPD decomposition on fair HYPERPLANE streams with $\rho_{\mathrm{CI}} = 0.05$ over 15 online classifiers. The first column shows $P(\hat{Y} = 1 \mid S = 0)$ and $P(\hat{Y} = 1 \mid S = 1)$; the second column shows $\Delta(\mathrm{TP})$ and $\Delta(\mathrm{FP})$, whose absolute sum gives the SPD profile.

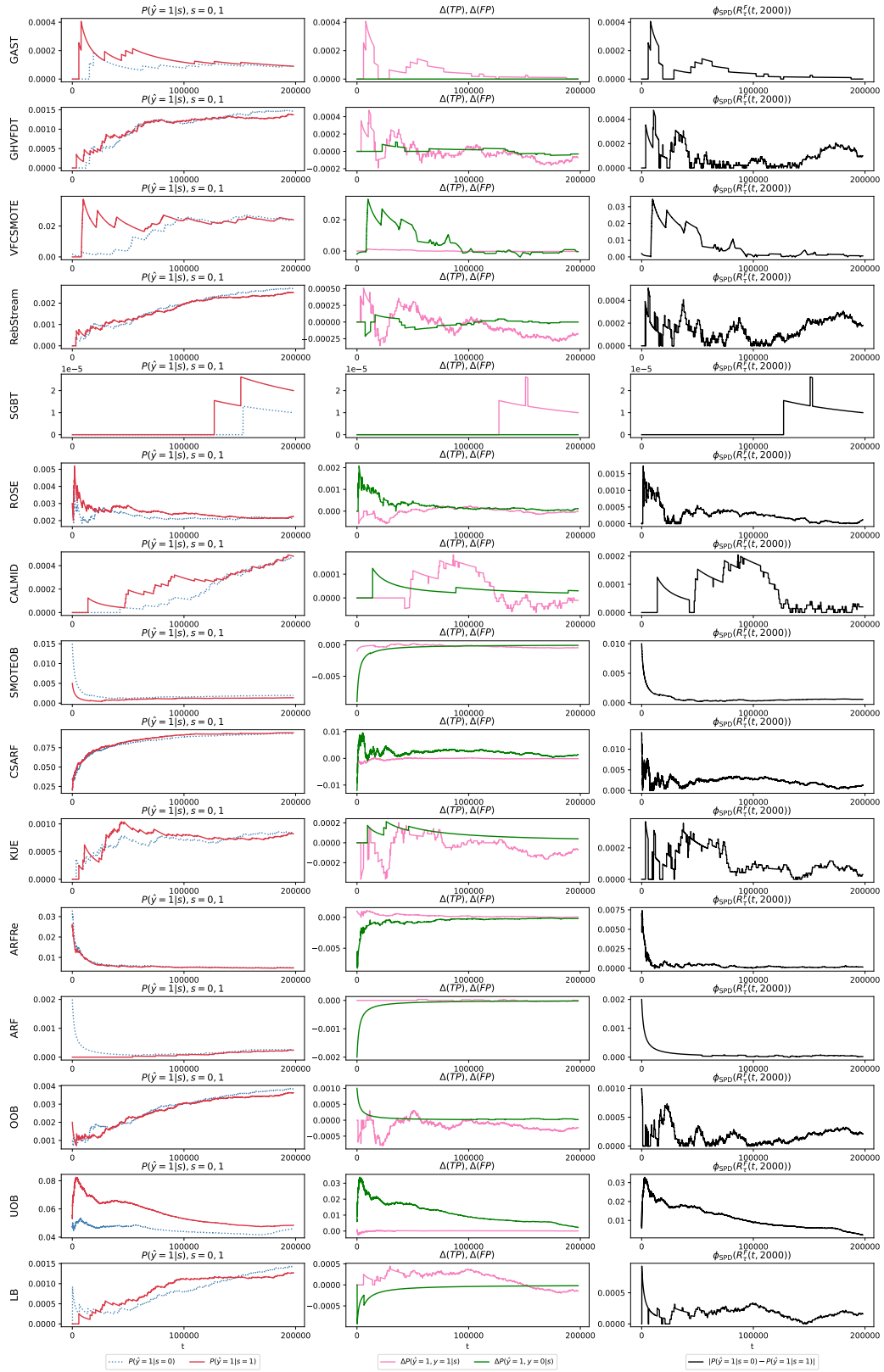


Figure 13: SPD decomposition on fair HYPERPLANE streams with $\rho_{\mathrm{CI}} = 0.01$ over 15 online classifiers. The first column shows $P(\hat{Y}=1 \mid S=0)$ and $P(\hat{Y}=1 \mid S=1)$; the second column shows $\Delta(\mathrm{TP})$ and $\Delta(\mathrm{FP})$, whose absolute sum gives the SPD profile.

F.4 Detailed evidence for Finding 2: SI broadly degrades group-based fairness in algorithmic decisions.

We next provide detailed evidence for the SI results. We use fair Hyperplane and RandomRBF streams, keep the class distribution balanced, and vary only the sensitive-group imbalance degree ρ_{SI}. In these streams, $S = 1$ is the minority sensitive group when $\rho_{\mathrm{SI}} < 0.50$. We evaluate the same 15 online stream classifiers used in the motivating analysis and use the same profile-based evaluation protocol as in the previous subsection. In the diagnostic panels, z denotes the sensitive attribute S.

Figures 14 and 15 show the overall SPD and DI results under different SI degrees. Unlike CI, where SPD and DI can show different sensitivities because positive prediction rates shrink under class scarcity, SI affects the two metrics in a more consistent direction. Across stationary streams with stronger SI, both SPD and DI generally increase for most algorithms. This indicates that sensitive-group imbalance directly affects group-wise prediction parity, rather than only changing the scale of positive predictions.

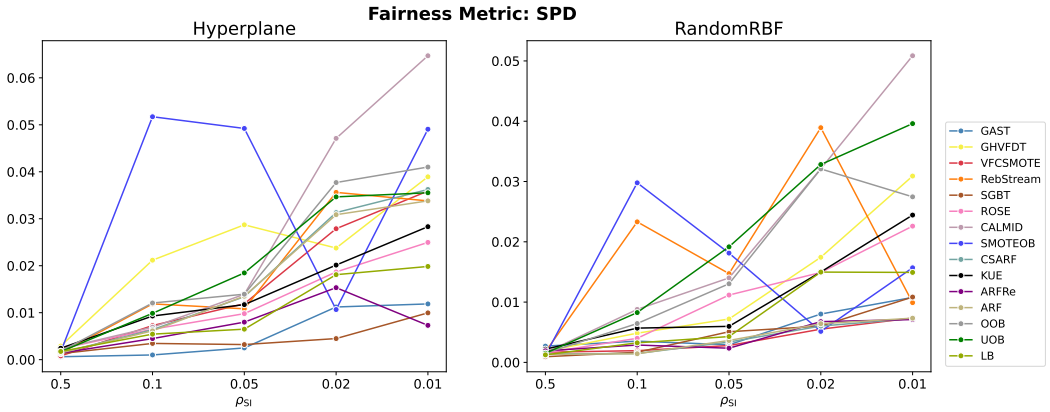


Figure 14: SPD of 15 online classifiers under fair Hyperplane and RandomRBF streams with different SI degrees.

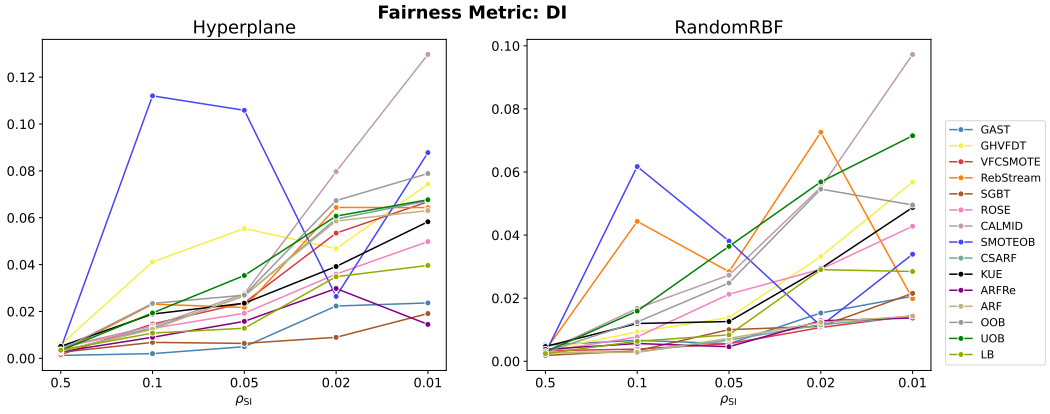


Figure 15: DI of 15 online classifiers under fair Hyperplane and RandomRBF streams with different SI degrees.

To better distinguish group-size effects on prediction volume from group-conditional prediction disparity, Figures 16–18 compare positive prediction rates (left), common-reference TP/FP count contrasts (middle), and SPD (right) for $\rho_{\mathrm{SI}} \in \{0.50, 0.05, 0.01\}$. Negative count contrasts can reflect fewer minority-group observations without implying worse fairness. The rate-based panels provide a complementary assessment of prediction disparity and minority-group fluctuations.

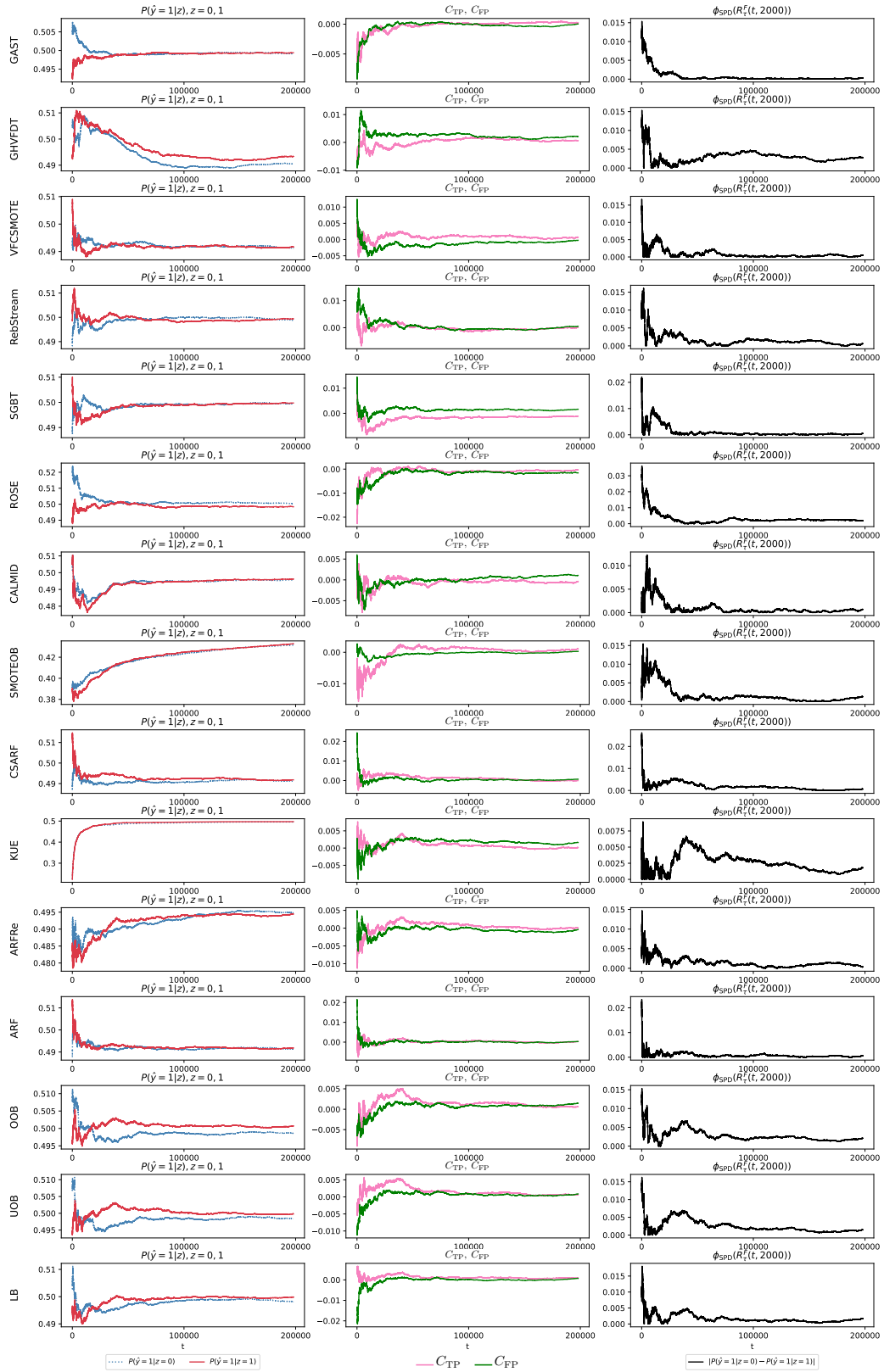


Figure 16: Prediction diagnostics on fair HYPERPLANE streams with $\rho_{\mathrm{SI}} = 0.50$ over 15 online classifiers. Left: group-conditional positive prediction rates. Middle: TP/FP count contrasts normalized by $N_{\tau,0}$. Right: SPD. In the panels, z denotes the sensitive attribute S.

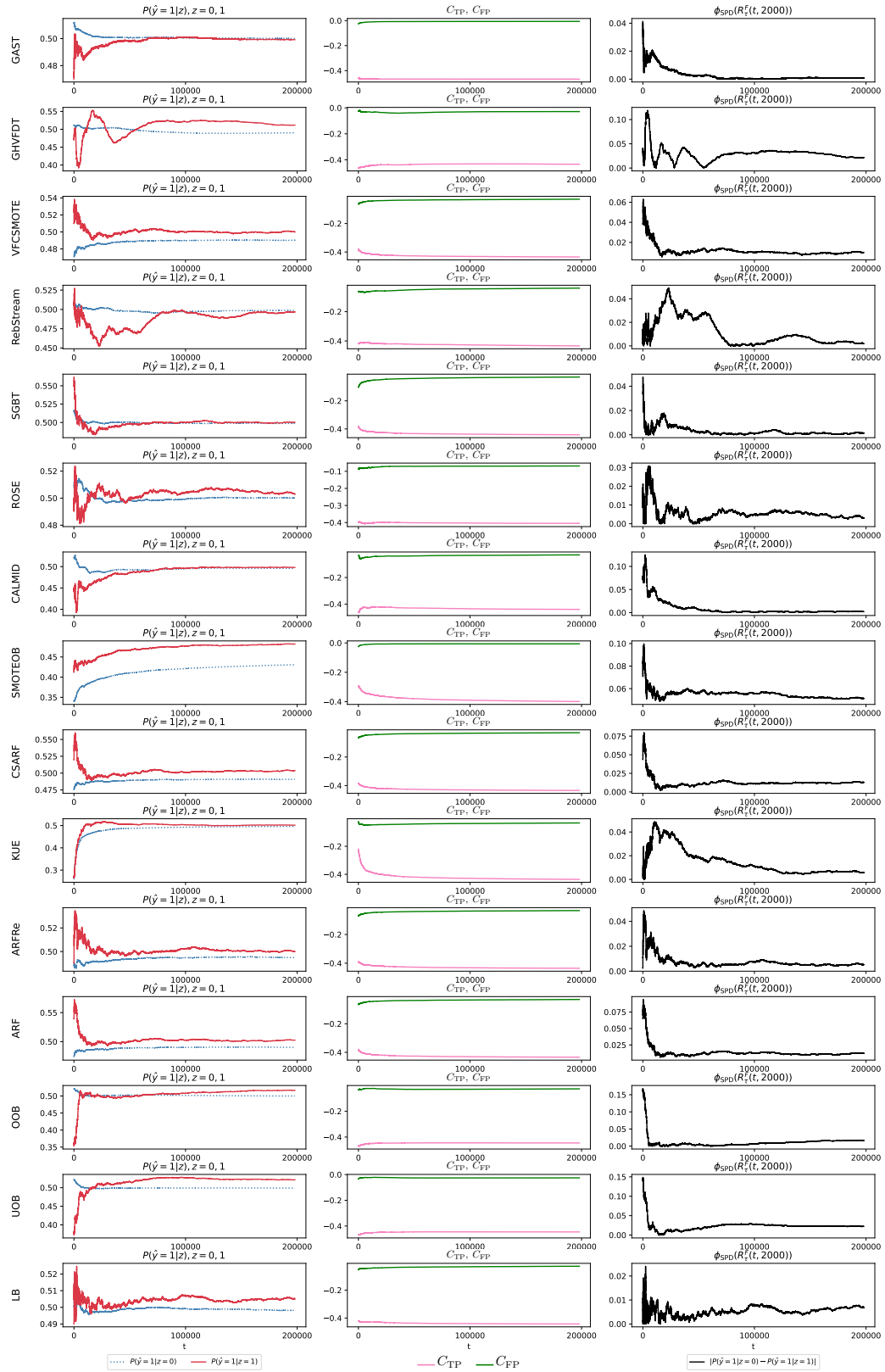


Figure 17: Prediction diagnostics on fair HYPERPLANE streams with $\rho_{\mathrm{SI}} = 0.05$ over 15 online classifiers. Left: group-conditional positive prediction rates. Middle: TP/FP count contrasts normalized by $N_{\tau,0}$. Right: SPD. In the panels, z denotes the sensitive attribute S.

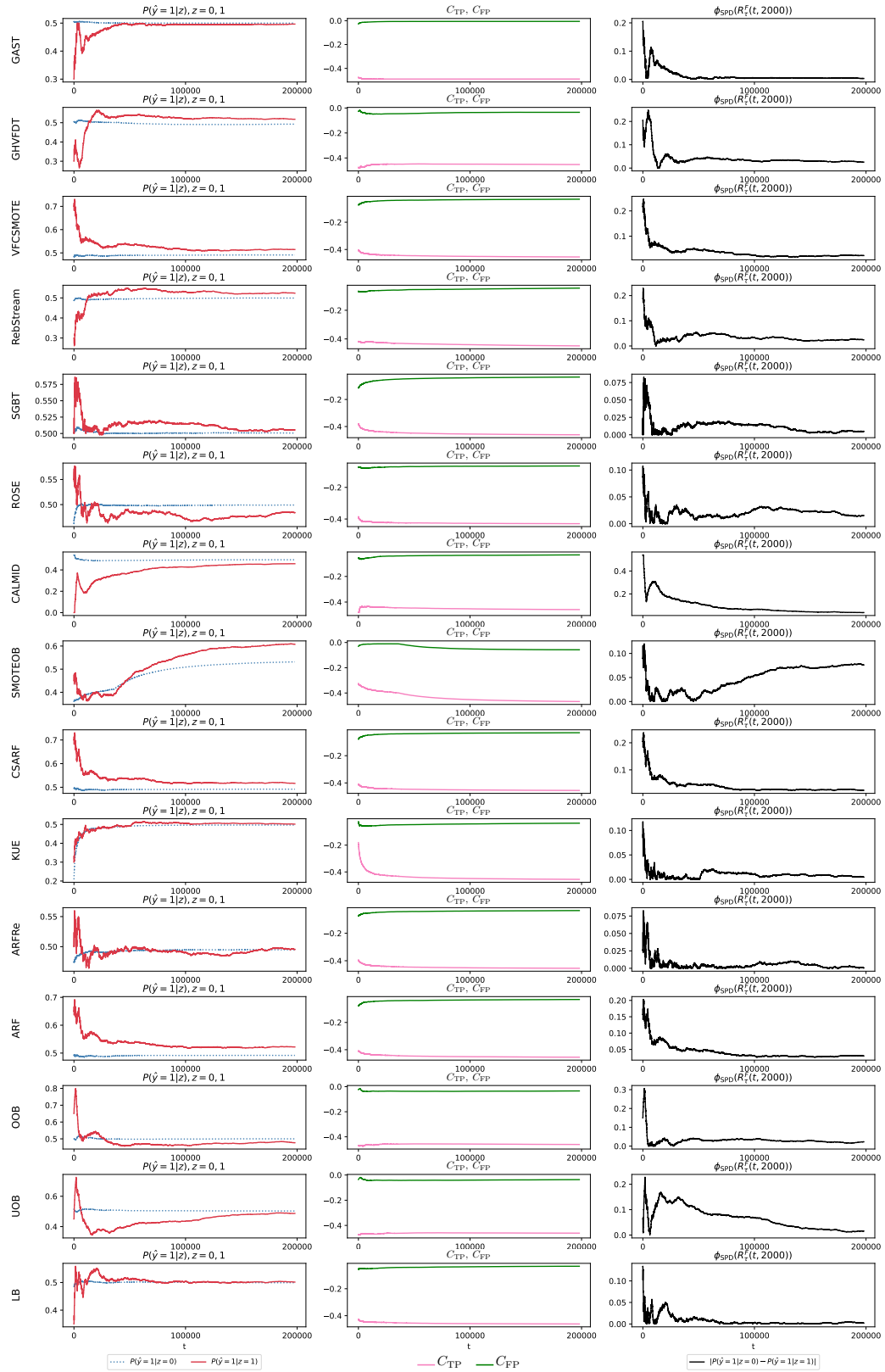


Figure 18: Prediction diagnostics on fair HYPERPLANE streams with $\rho_{\mathrm{SI}} = 0.01$ over 15 online classifiers. Left: group-conditional positive prediction rates. Middle: TP/FP count contrasts normalized by $N_{\tau,0}$. Right: SPD. In the panels, z denotes the sensitive attribute S.

F.5 Detailed evidence for Finding 3: Imbalance amplifies algorithmic unfairness inherited from biased data streams.

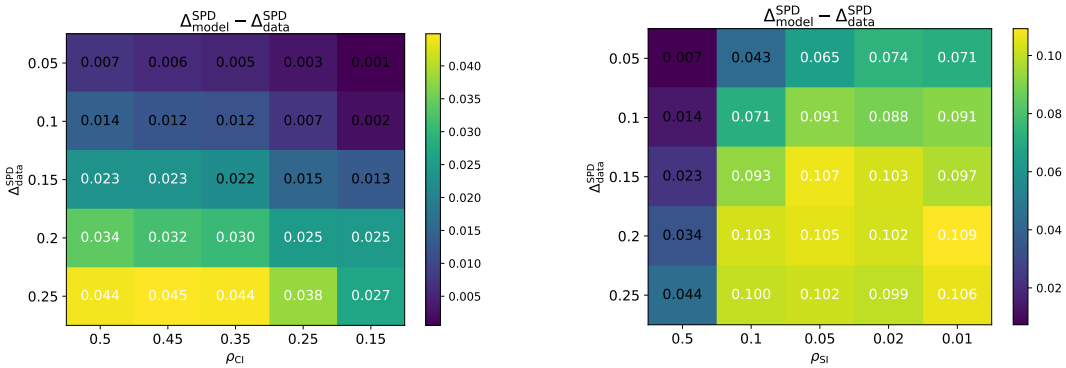
We further analyze how imbalance interacts with data-side unfairness. For each stream, we compare the model-side SPD gap with the label-side SPD gap:

$$\Delta^{\mathrm{SPD}}_{\mathrm{amp}} = \Delta^{\mathrm{SPD}}_{\mathrm{model}} - \Delta^{\mathrm{SPD}}_{\mathrm{data}}, \tag{64}$$

where $\Delta^{\mathrm{SPD}}_{\mathrm{model}} = |P(\hat{Y} = 1 \mid S = 0) - P(\hat{Y} = 1 \mid S = 1)|$ and $\Delta^{\mathrm{SPD}}_{\mathrm{data}} = |P(Y = 1 \mid S = 0) - P(Y = 1 \mid S = 1)|$. A positive value means that the learned model exhibits a larger group disparity than the disparity already present in the stream labels. We compute $\Delta^{\mathrm{SPD}}_{\mathrm{amp}}$ for each of the 15 algorithms in the motivating analysis and report the average value in the heatmaps.

Figure 19 shows the amplification patterns under CI and SI. Under CI, the average values are mostly positive, indicating that online learners can amplify the unfairness inherited from biased data streams. However, for a fixed $\Delta^{\mathrm{SPD}}_{\mathrm{data}}$, the amplification becomes weaker in streams with stronger CI. This is consistent with the previous CI analysis: severe class imbalance reduces the overall positive prediction level, which can also reduce the absolute SPD gap.

The SI pattern is different. For a fixed $\Delta^{\mathrm{SPD}}_{\mathrm{data}}$, stronger SI generally leads to larger amplification. For a fixed SI degree, larger data-side unfairness also leads to stronger amplification. This suggests that sensitive-group imbalance makes online learners more likely to strengthen the group disparity already present in the stream distribution.

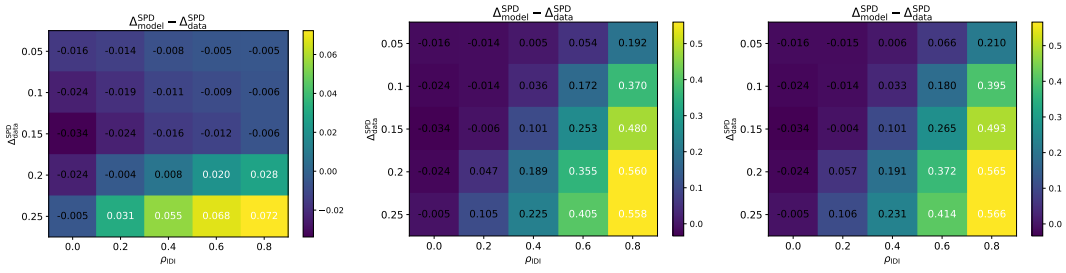


(a) CI (b) SI

Figure 19: Amplification of data-side unfairness under CI and SI. Each cell reports the average $\Delta^{\mathrm{SPD}}_{\mathrm{model}} - \Delta^{\mathrm{SPD}}_{\mathrm{data}}$ over 15 online classifiers. Rows correspond to different data-side unfairness levels $\Delta^{\mathrm{SPD}}_{\mathrm{data}}$, and columns correspond to different imbalance degrees. Positive values indicate that the model-side SPD is larger than the label-side SPD.

Figure 20 reports the corresponding results under IDI. The borderline-dominated setting shows relatively weak amplification compared with the rare- and outlier-dominated settings. In contrast, rare and outlier dominance lead to clearer amplification when ρ_{IDI} is large. For a fixed $\Delta^{\mathrm{SPD}}_{\mathrm{data}}$, stronger rare or outlier dominance usually increases $\Delta^{\mathrm{SPD}}_{\mathrm{amp}}$. For a fixed IDI degree, larger data-side unfairness also tends to produce larger amplification.

These results suggest that rare and outlier instances make the inherited data-side disparity harder to control during online updates. Borderline instances mainly affect samples near the decision boundary and do not show the same level of amplification. This supports the main observation that IDI-induced fairness degradation is driven more by rare and outlier dominance than by borderline dominance.



(a) Borderline (b) Rare (c) Outlier

Figure 20: Amplification of data-side unfairness under IDI. Each cell reports the average $\Delta^{\mathrm{SPD}}_{\mathrm{model}} - \Delta^{\mathrm{SPD}}_{\mathrm{data}}$ over 15 online classifiers. Rows correspond to different data-side unfairness levels $\Delta^{\mathrm{SPD}}_{\mathrm{data}}$, and columns correspond to different IDI degrees. Positive values indicate that the model amplifies the label-side SPD gap.

F.6 Detailed evidence for Finding 4: IDI severity amplifies fairness degradation, especially under rare and outlier dominance.

We further examine IDI using fair Difficulty streams. The CI and SI are kept balanced, while the stream composition is changed by increasing the proportion of one difficult instance type. We consider borderline-, rare-, and outlier-dominated settings with $\rho_{\mathrm{IDI}} \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$, where $\rho_{\mathrm{IDI}} = 0.0$ corresponds to the safe-only baseline. The same 15 online stream classifiers are evaluated under the shared prequential protocol.

Figures 21 and 22 summarize the overall SPD and DI trends. The borderline setting has a relatively mild effect on most algorithms. In contrast, rare- and outlier-dominated streams lead to larger SPD and DI for several classifiers, especially at larger ρ_{IDI}. This indicates that IDI can degrade algorithmic fairness, and that this degradation is more evident when rare or outlier instances dominate the stream.

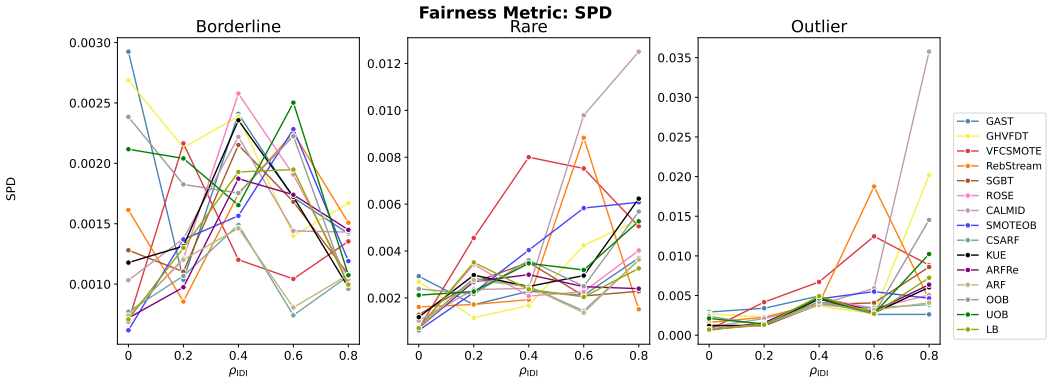


Figure 21: SPD of 15 online classifiers under fair Difficulty streams with different IDI degrees.

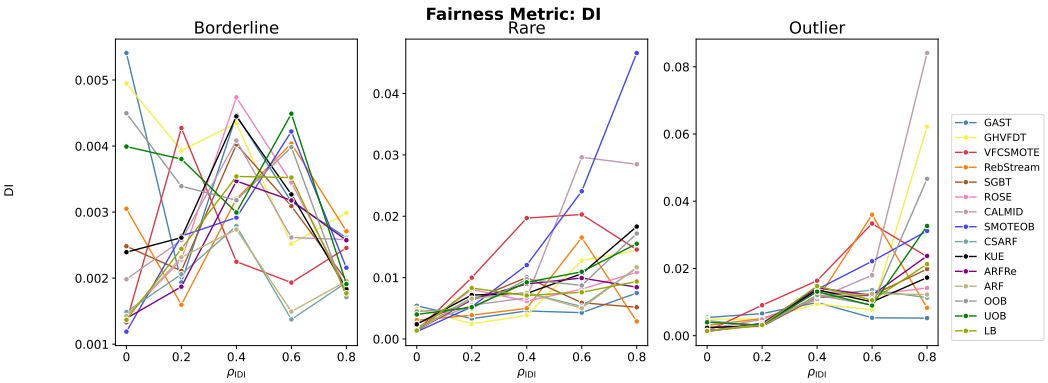


Figure 22: DI of 15 online classifiers under fair DIFFICULTY streams with different IDI degrees.

Figures 23–29 provide the corresponding profile-level diagnostics. The layout follows the CI decomposition figures: the first column shows group-wise positive prediction rates, the second column shows the two SPD decomposition terms, and the third column gives the SPD profile. The safe-only stream in Figure 23 serves as the reference. Compared with this reference, the borderline profiles in Figures 24 and 25 show relatively limited changes in the SPD profile for most algorithms.

The rare and outlier profiles show a more visible effect. In Figures 26–27 and Figures 28–29, the two decomposition terms fluctuate more noticeably for several algorithms, and the resulting SPD profiles are often larger than in the borderline setting. A plausible explanation is that rare and outlier instances provide sparser or less stable local evidence, making group-wise prediction behavior harder to align. This observation is consistent with the main-text finding that fairness degradation under IDI is mainly driven by rare- and outlier-dominated streams.

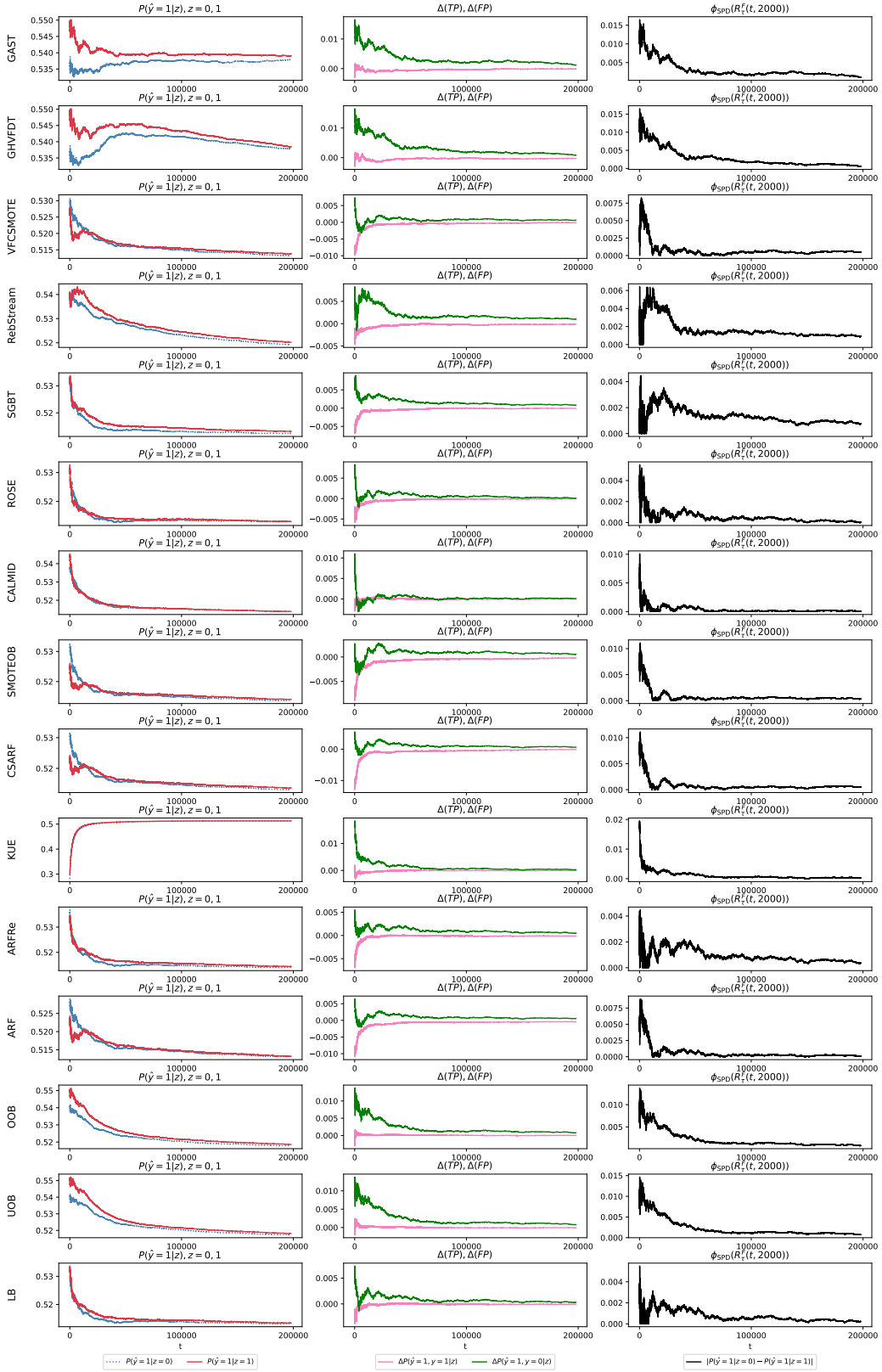


Figure 23: Prediction-rate and SPD profiles on fair DIFFICULTY streams with $\rho_{\text{IDI}} = 0.0$ over 15 online classifiers. The layout follows the CI decomposition figures; s denotes the sensitive attribute.

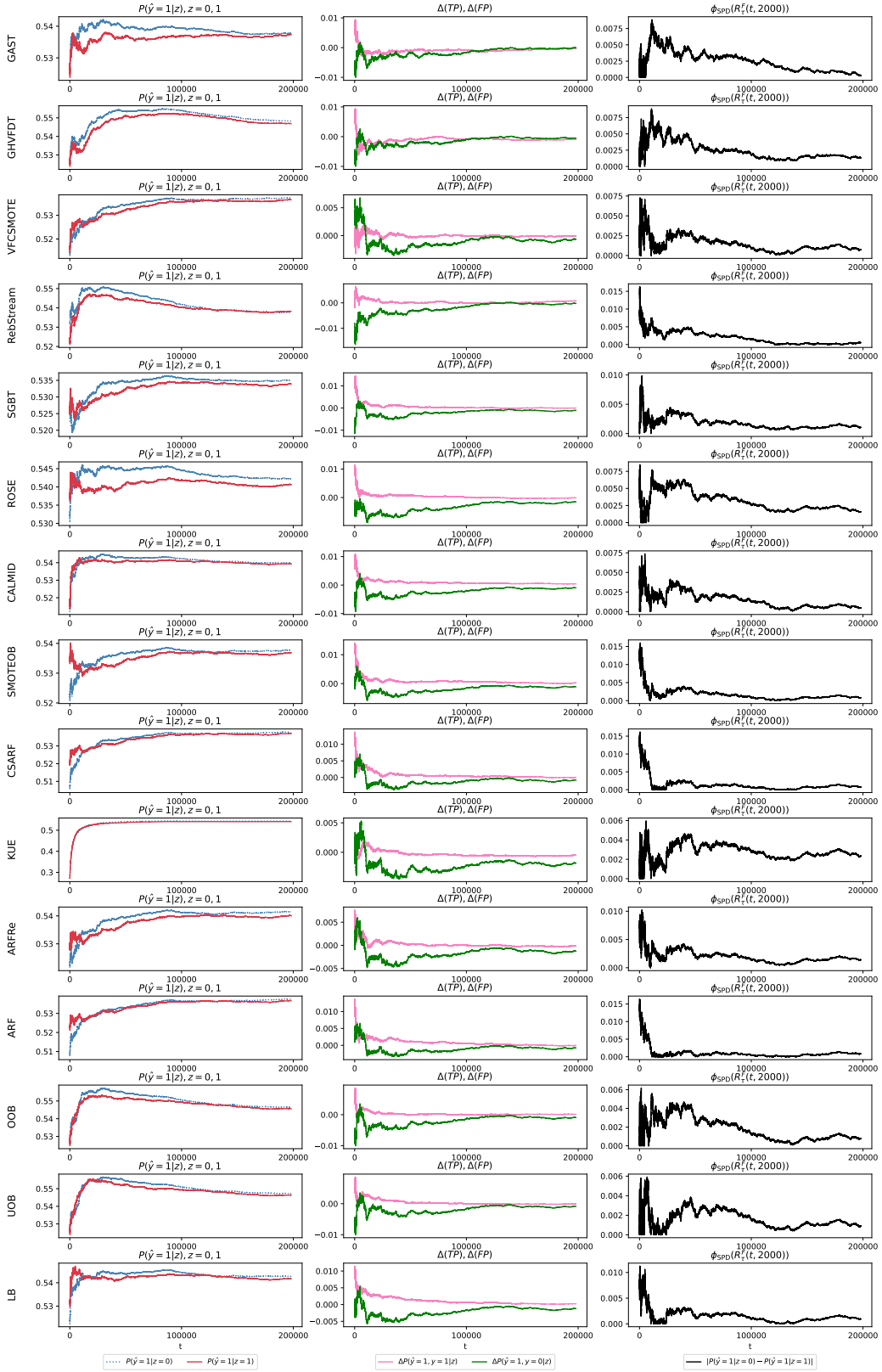


Figure 24: Prediction-rate and SPD profiles on fair DIFFICULTY streams with $\rho_{\text{IDI}} = 0.4$ over 15 online classifiers. The layout follows the CI decomposition figures; s denotes the sensitive attribute.

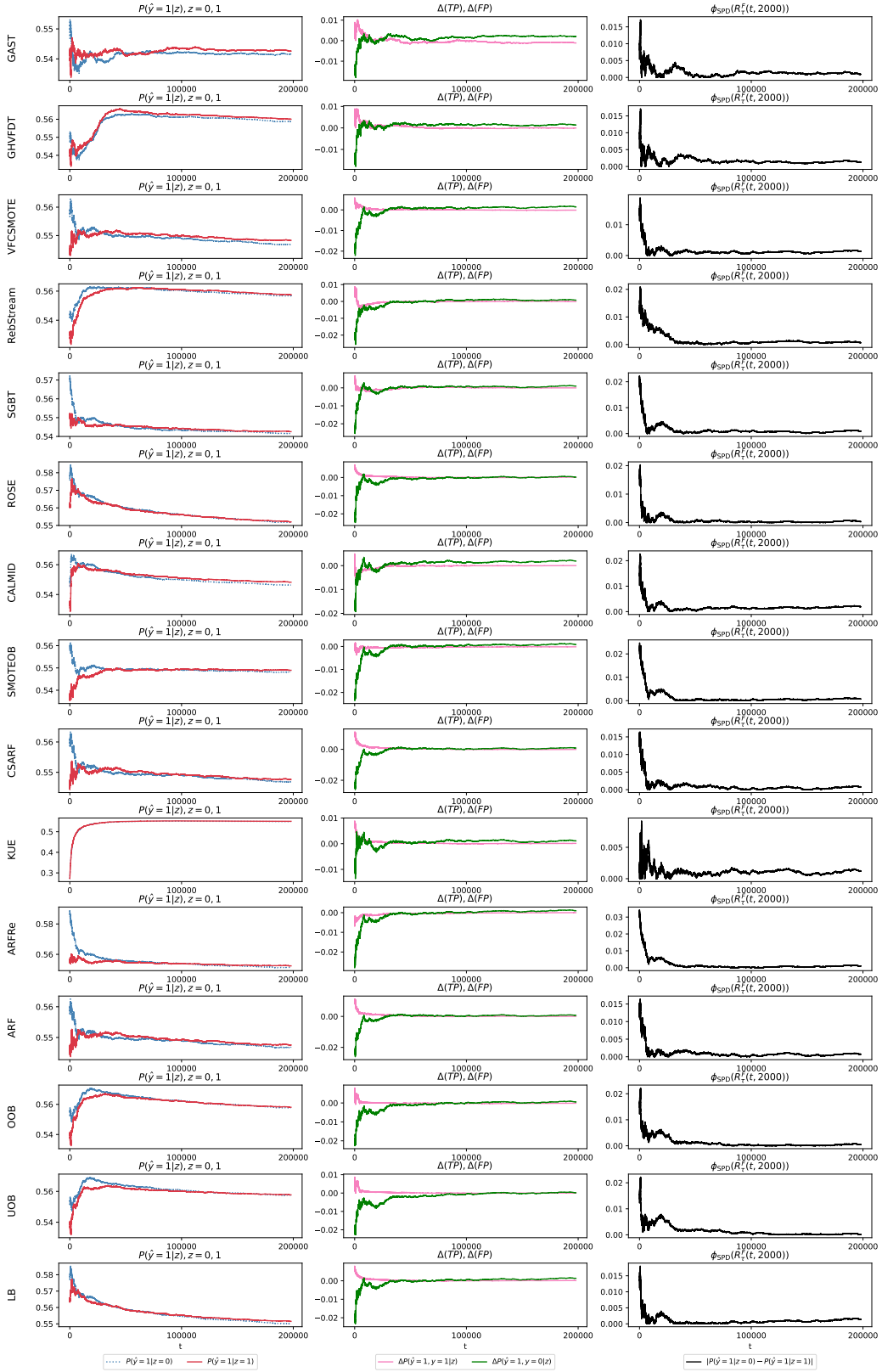


Figure 25: Prediction-rate and SPD profiles on fair DIFFICULTY streams with $\rho_{\text{IDI}} = 0.8$ over 15 online classifiers. The layout follows the CI decomposition figures; s denotes the sensitive attribute.

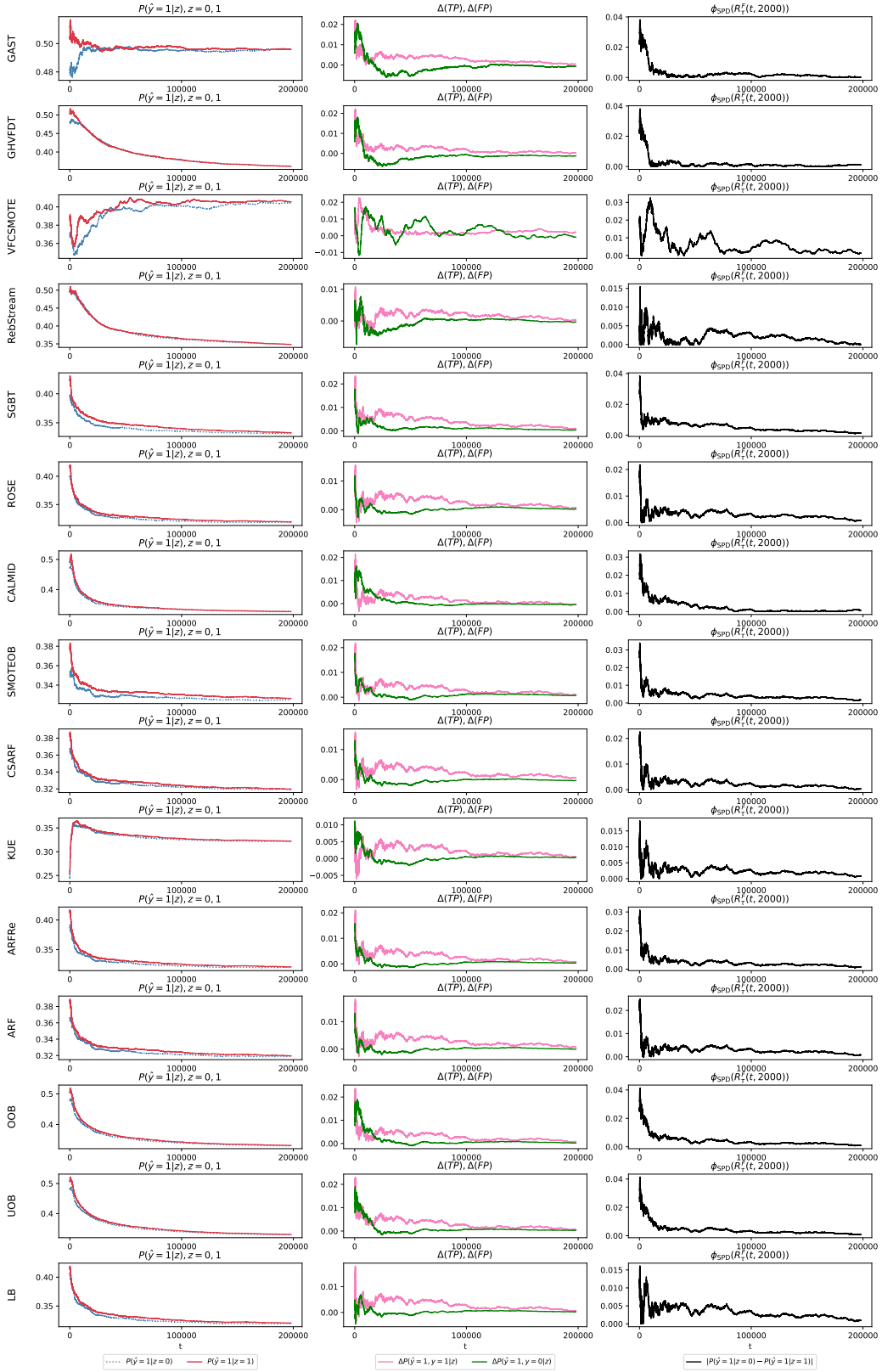


Figure 26: Prediction-rate and SPD profiles on fair DIFFICULTY streams with $\rho_{\mathrm{IDI}} = 0.4$ over 15 online classifiers. The layout follows the CI decomposition figures; s denotes the sensitive attribute.

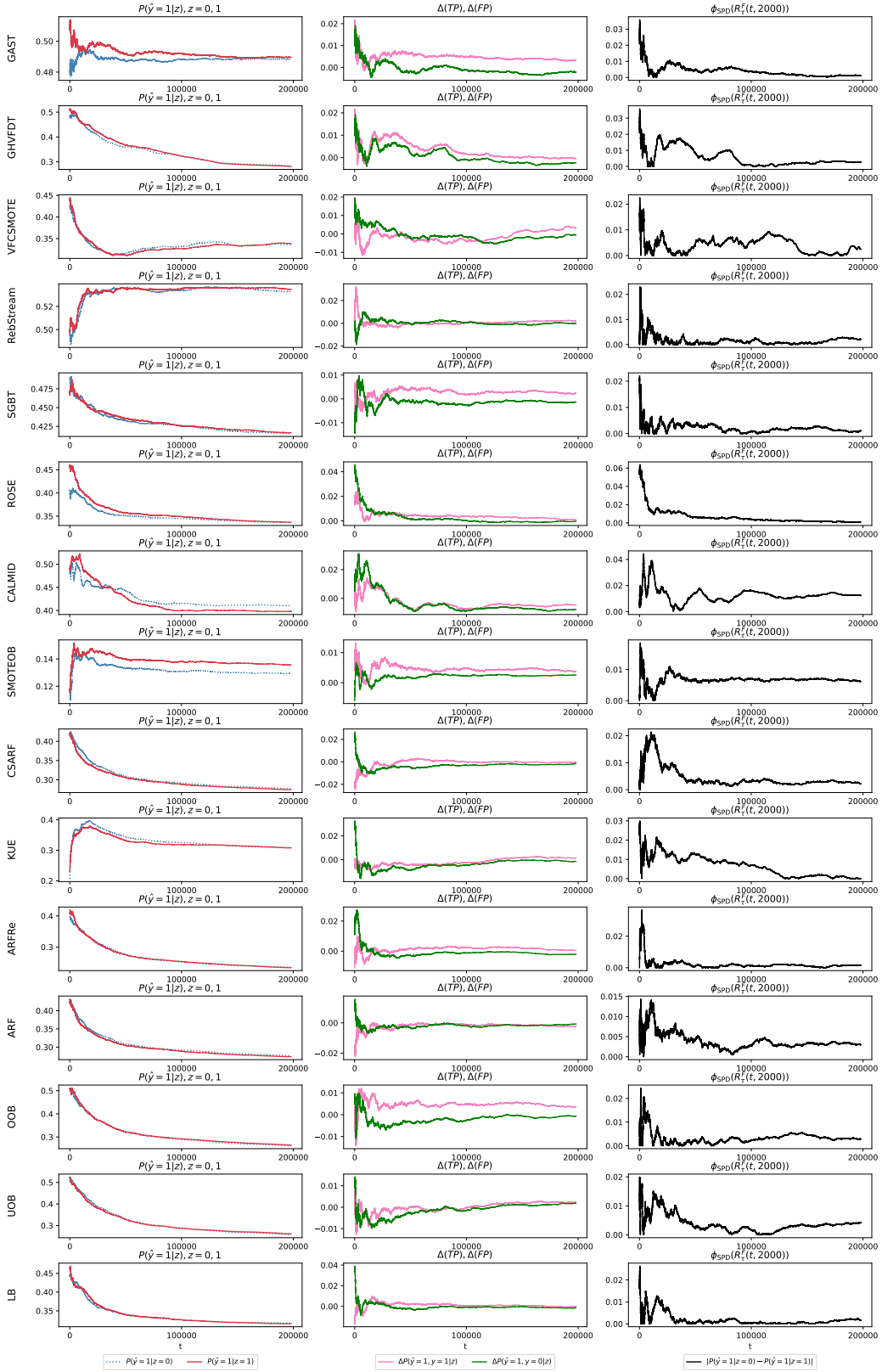


Figure 27: Prediction-rate and SPD profiles on fair DIFFICULTY streams with $\rho_{\mathrm{IDI}} = 0.8$ over 15 online classifiers. The layout follows the CI decomposition figures; s denotes the sensitive attribute.

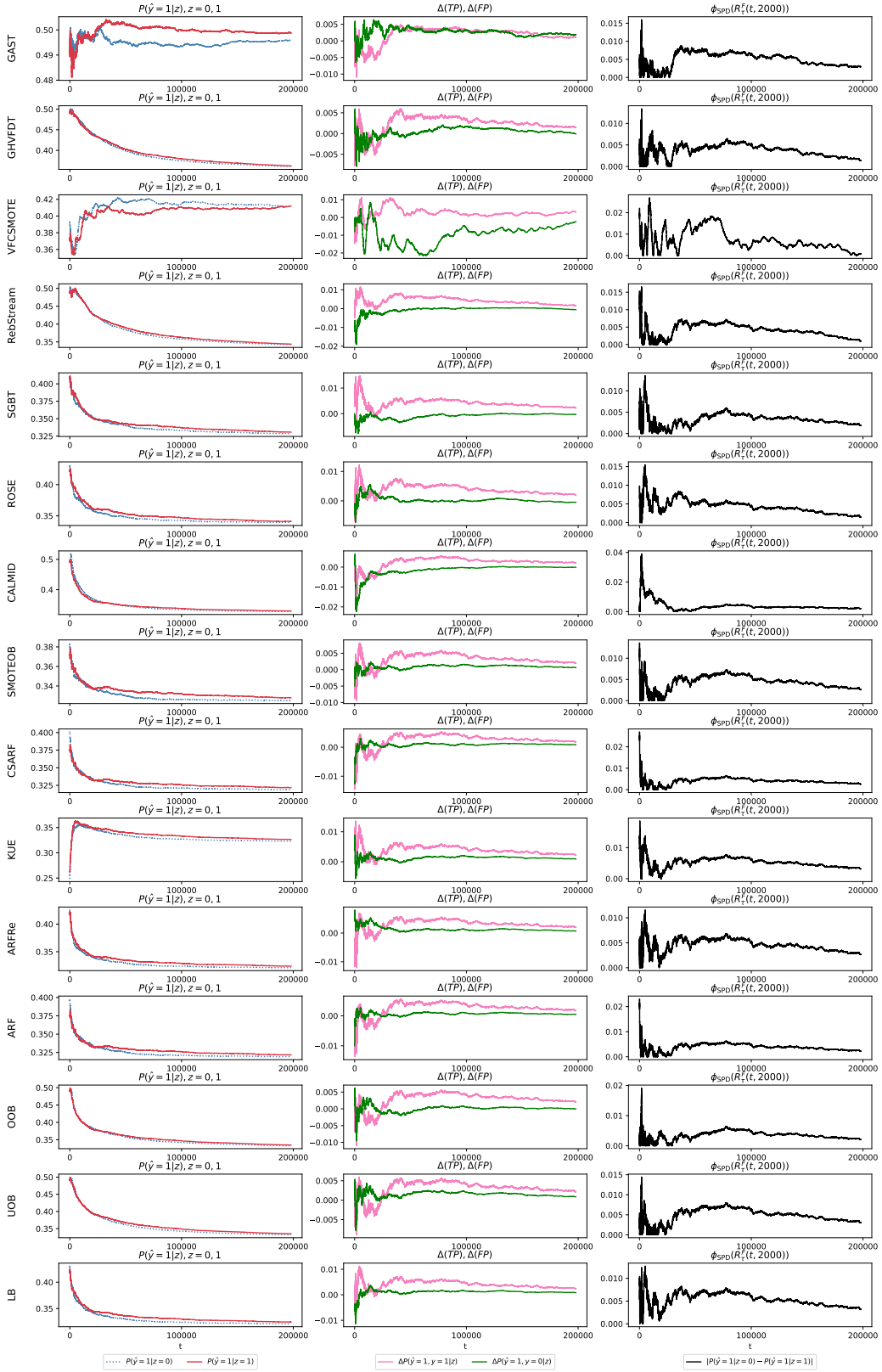


Figure 28: Prediction-rate and SPD profiles on fair DIFFICULTY streams with $\rho_{\mathrm{IDI}} = 0.4$ over 15 online classifiers. The layout follows the CI decomposition figures; s denotes the sensitive attribute.

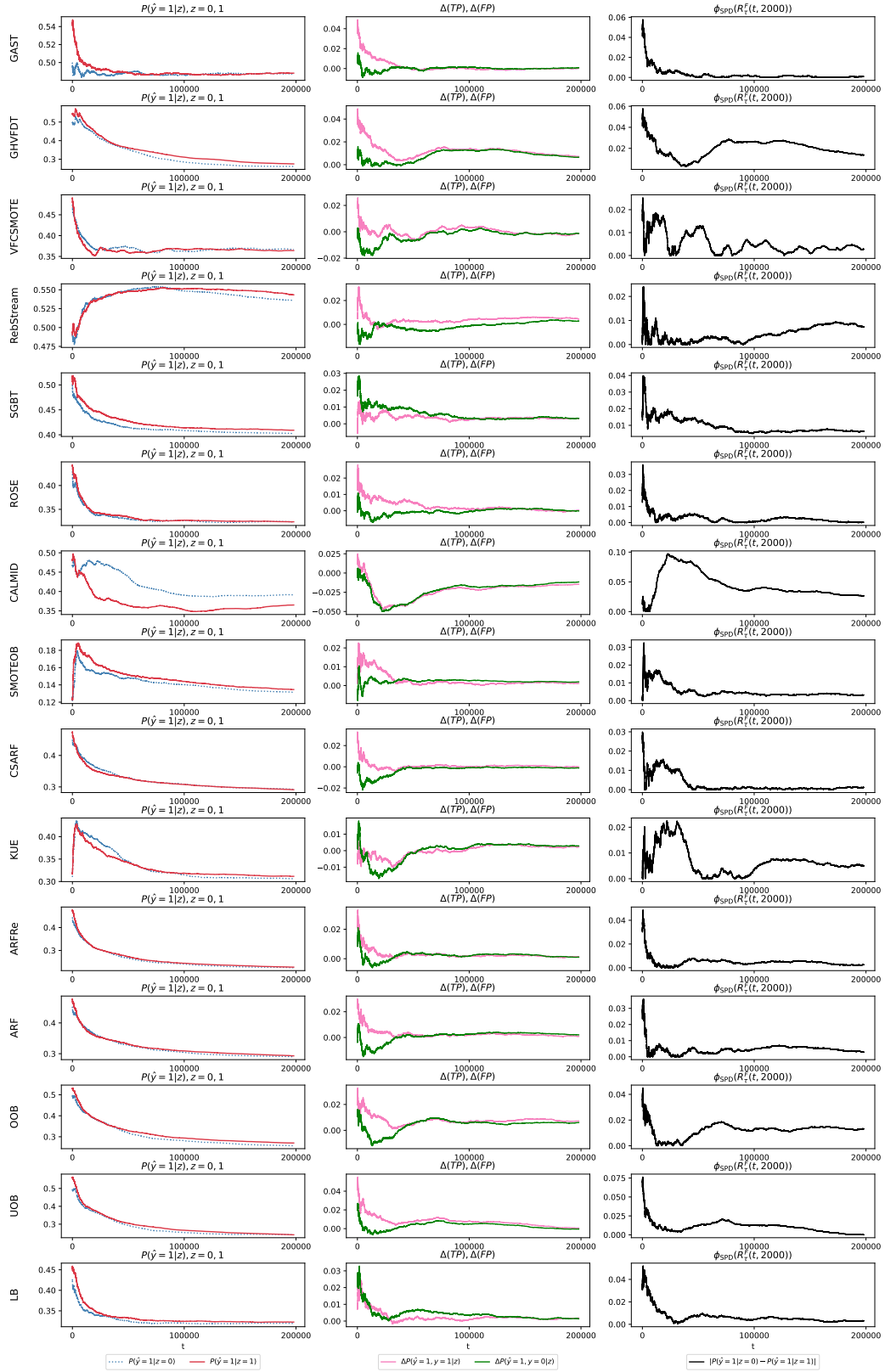
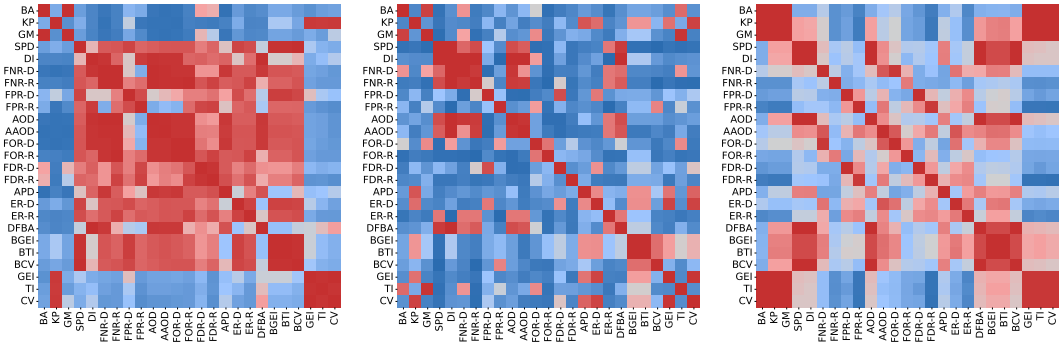


Figure 29: Prediction-rate and SPD profiles on fair DIFFICULTY streams with $\rho_{\mathrm{IDI}} = 0.8$ over 15 online classifiers. The layout follows the CI decomposition figures; s denotes the sensitive attribute.

F.7 Additional Finding: Imbalance sources alter agreement among algorithmic performance metrics.

Figure 30 shows that imbalance settings can substantially affect metric agreement. Although all panels use the same metrics, their Spearman correlations across stationary streams change markedly. Comparing only CI in Figure 30a with only SI in Figure 30c, even under fair stream data, the correlation structures are clearly different. This suggests that CI and SI reshape the relations among algorithmic fairness metrics in different ways. Within CI, stream-data unfairness further changes metric agreement. Under fair stream data in Figure 30a, several fairness metrics form a positive-correlation block, such as FOR- and FNR-related metrics. However, when the stream data are unfair in Figure 30b, this block becomes fragmented and some associations even reverse, indicating potentially conflicting algorithmic fairness rankings. Accuracy metrics are also affected; since G-Mean largely overlaps with either BA or KP in later comparisons, we report BA and KP as primary accuracy metrics in our experimental comparisons.



(a) Only CI, $\Delta^{\mathrm{SPD}}_{\mathrm{data}} = 0$ (b) Only CI, $\Delta^{\mathrm{SPD}}_{\mathrm{data}} = 0.25$ (c) Only SI, $\Delta^{\mathrm{SPD}}_{\mathrm{data}} = 0$

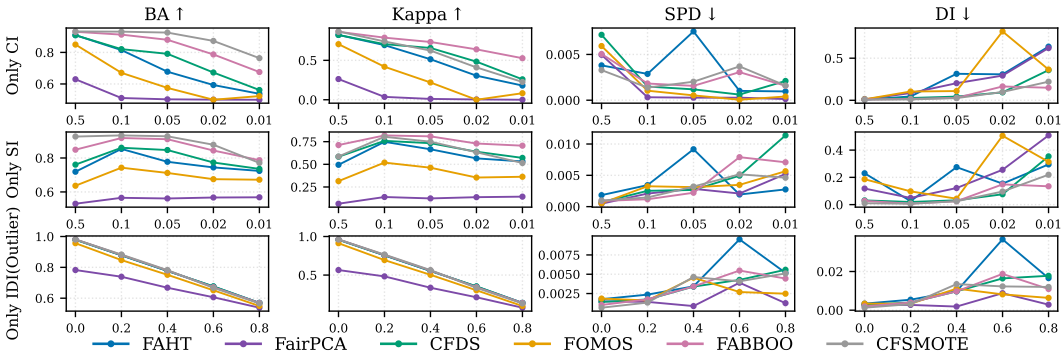
Figure 30: Imbalance settings reshape metric agreement. Each panel reports Spearman correlations among accuracy and algorithmic fairness metrics across stationary streams under CI/SI and fair/unfair data. Red and blue denote positive and negative correlations; intensity indicates strength.

G Additional Results for RQ1

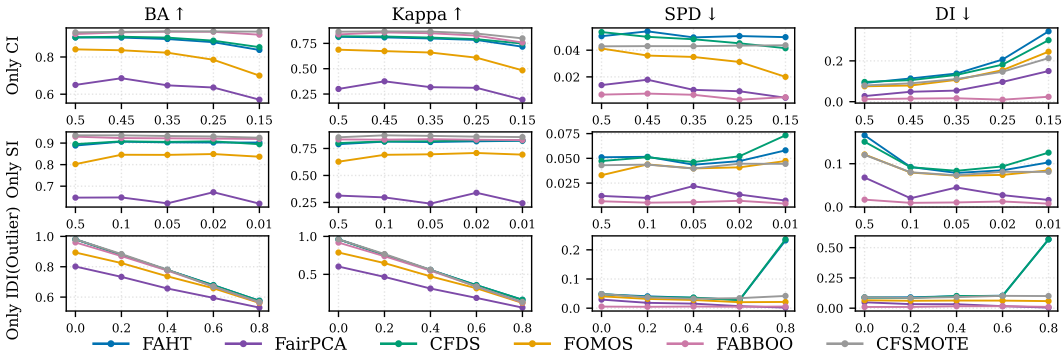
G.1 Detailed evidence for Findings 5–7

We provide the full RQ1 results under different data-side unfairness levels. Specifically, Figures 31 and 32 report the performance of six fairness-aware streaming methods when $\Delta^{\mathrm{SPD}}_{\mathrm{data}} \in \{0, 0.05, 0.15, 0.20, 0.25\}$. For each unfairness level, we vary one imbalance source at a time, including CI, SI, and outlier-dominated IDI, and report BA, Kappa, SPD, and DI.

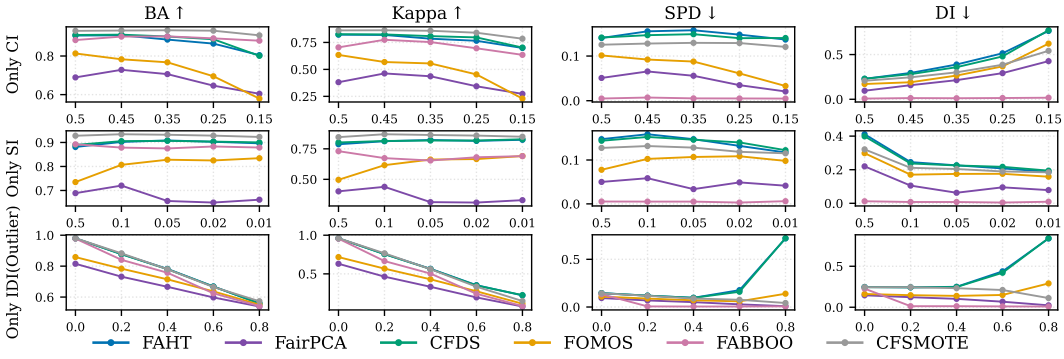
The results are consistent with the main-text findings. Methods without explicit imbalance modeling do not show stable fairness improvement across all imbalance sources and data-side unfairness levels. Their SPD and DI can fluctuate under stronger CI or SI, and the instability becomes more visible under outlier-dominated IDI. Methods with stronger parity control, especially FABBOO, often obtain low SPD and DI, but this is accompanied by lower BA and Kappa in many settings. CFSMOTE usually preserves stronger predictive performance, but its SPD and DI remain sensitive under IDI. Overall, these results support the main conclusion that fairness-aware streaming methods should account for imbalance during online learning, and that group-level rebalancing alone is not sufficient when instance-level difficulty dominates the stream.



(a) $\Delta_{\text{data}}^{\text{SPD}} = 0$

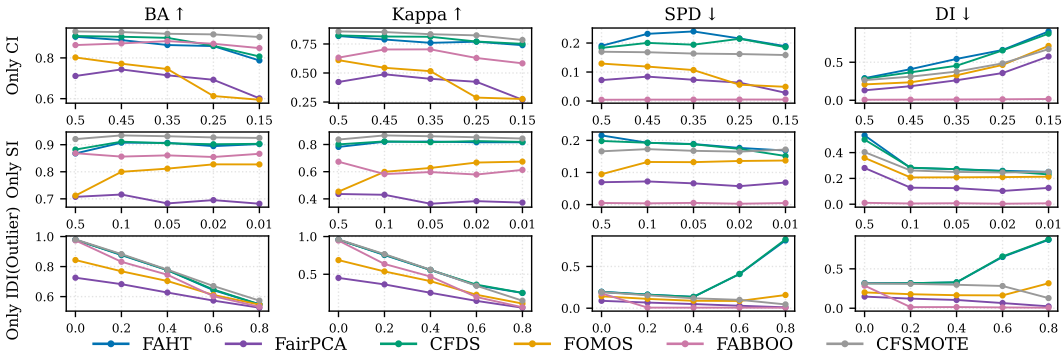


(b) $\Delta_{\text{data}}^{\text{SPD}} = 0.05$

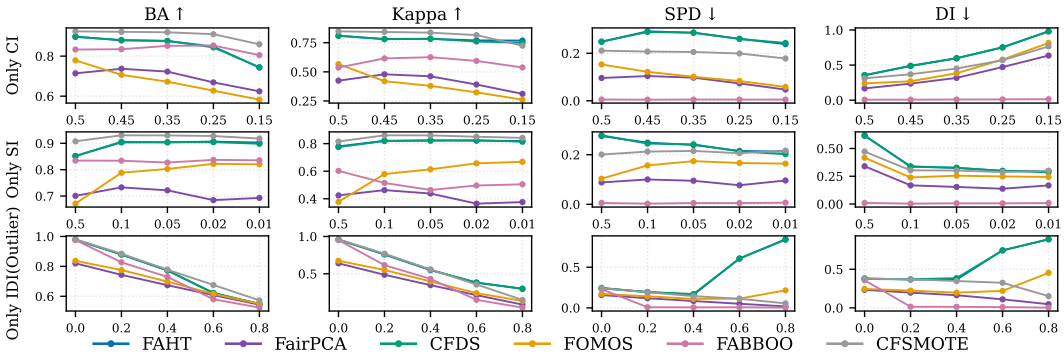


(c) $\Delta_{\text{data}}^{\text{SPD}} = 0.15$

Figure 31: Detailed RQ1 results under fair and mildly unfair stream data. Each subfigure reports BA, Kappa, SPD, and DI for six fairness-aware streaming methods under CI, SI, and outlier-dominated IDI. Rows vary one imbalance source at a time; columns report evaluation metrics.



(a) $\Delta_{\text{data}}^{\text{SPD}} = 0.20$



(b) $\Delta_{\text{data}}^{\text{SPD}} = 0.25$

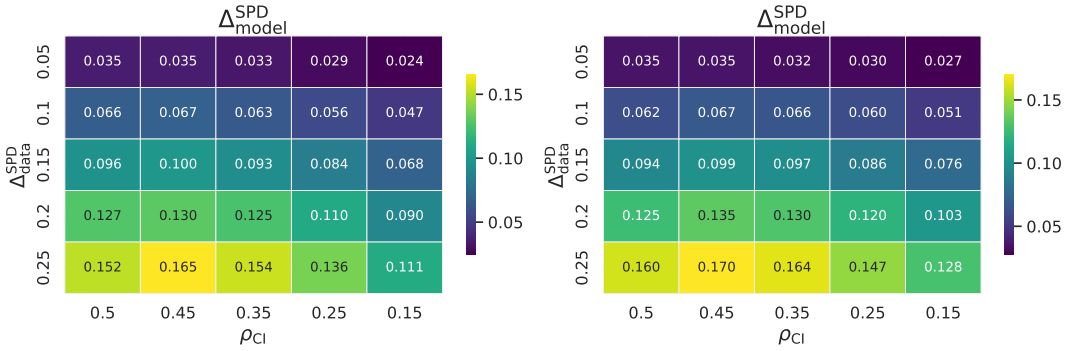
Figure 32: Detailed RQ1 results under increasingly unfair stream data. Each subfigure reports BA, Kappa, SPD, and DI for six fairness-aware streaming methods under CI, SI, and outlier-dominated IDI. Rows vary one imbalance source at a time; columns report evaluation metrics.

G.2 Additional Finding: Existing methods only partially reduce inherited data-side disparity.

The main text analyzes the stability and accuracy–fairness trade-offs of existing fairness-aware stream methods. Here we further ask whether these methods can suppress model-side disparity when the stream labels already contain group-dependent bias. For each setting, we vary $\Delta_{\text{data}}^{\text{SPD}}$ and one imbalance source, and report the average model-side SPD over the compared fairness-aware stream methods.

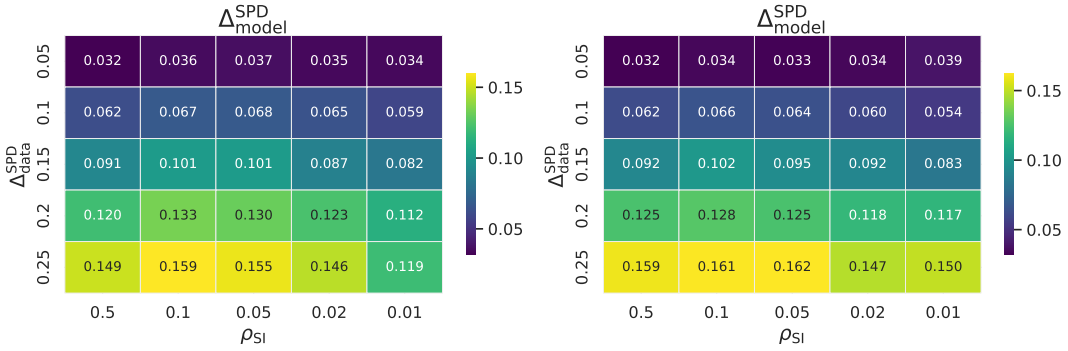
Figures 33–35 show that existing methods only partially reduce the disparity inherited from biased stream data. Across CI, SI, and IDI settings, larger $\Delta_{\text{data}}^{\text{SPD}}$ generally corresponds to larger model-side SPD. This suggests that, although these methods can improve fairness in some settings, they do not fully decouple online predictions from the label–group disparity already present in the stream.

The difficulty becomes more pronounced when data-side bias is combined with imbalance. Under CI and SI, model-side SPD remains non-negligible across both HYPERPLANE and RANDOMRBF streams. Under IDI, rare- and outlier-dominated streams lead to larger residual disparity than borderline-dominated streams in several settings. This provides an additional reason why biased and imbalanced streams remain challenging for existing fairness-aware online methods.



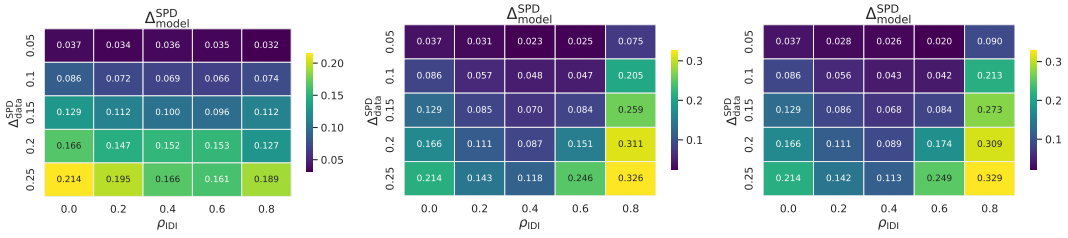
(a) HYPERPLANE (b) RANDOMRBF

Figure 33: Model-side SPD under biased streams with different CI degrees. Rows correspond to $\Delta^{\text{SPD}}_{\text{data}}$, and columns correspond to ρ_{CI}. Each cell summarizes the model-side SPD of the compared fairness-aware streaming methods; smaller values indicate better fairness.



(a) HYPERPLANE (b) RANDOMRBF

Figure 34: Model-side SPD under biased streams with different SI degrees. Rows correspond to $\Delta^{\text{SPD}}_{\text{data}}$, and columns correspond to ρ_{SI}. Each cell summarizes the model-side SPD of the compared fairness-aware streaming methods; smaller values indicate better fairness.



(a) Borderline (b) Rare (c) Outlier

Figure 35: Model-side SPD under biased DIFFICULTY streams with different IDI degrees. Rows correspond to $\Delta^{\text{SPD}}_{\text{data}}$, and columns correspond to ρ_{IDI}. Each cell summarizes the model-side SPD of the compared fairness-aware streaming methods; smaller values indicate better fairness.

H WELL IMPLEMENTATION DETAILS

Instantiation of the diversification operator. In our implementation, we instantiate $\mathcal{D}_m$ by multiplicative mean-preserving perturbation. Specifically, we sample $\tilde{\xi}_t^{(m)} = \exp(\epsilon_t^{(m)})$ with

$\epsilon_t^{(m)} \sim \mathcal{N}(0, \sigma^2)$ and normalize

$$\xi_t^{(m)} = \frac{M\tilde{\xi}_t^{(m)}}{\sum_{j=1}^{M} \tilde{\xi}_t^{(j)}}.$$

The learner-specific update weight is then

$$\omega_t^{(m)} = \omega_t \xi_t^{(m)}.$$

This preserves the ensemble-average training influence of instance t while allowing different base learners to receive slightly different update strengths.

Online statistics and difficulty buffer. WELL maintains exponentially discounted statistics for the label–sensitive cell distribution and for the prediction-side fairness distribution. The forgetting factor θ is used for the recent label–sensitive cell masses that determine the group-imbalance component, while θ_{fair} is used for the contingency statistics over (Y, S) and $(\hat{Y}, S)$ that determine the distributional and decision-disparity components. The difficulty score is computed from a FIFO buffer of recent instances. Specifically, WELL stores recent feature vectors and labels in a buffer of size B, retrieves the nearest neighbors of the current instance, and uses their label agreement with y_t to obtain the local reliability score r_t and difficulty score $d_t = 1 - r_t$. The sensitive attribute is excluded from the distance computation when estimating local difficulty, so that the difficulty term reflects local feature–label consistency rather than group identity.

Prediction-disparity feedback and difficulty moderation. The decision-disparity strength is updated from recent prediction-side disparity. When the empirical prediction gap is small, the model-side correction remains weak; as the gap increases, the correction strength is increased but capped by $\gamma_{\text{dec}}^{\max}$ to avoid unstable updates. The distributional and decision-disparity coefficients are also bounded by $\alpha_{\max}$, which prevents the fairness-related factors from dominating the imbalance and difficulty terms. The difficulty moderation function $\chi(d_t)$ is implemented as a decreasing sigmoid-like penalty, where the slope is controlled by k_χ. Thus, larger k_χ makes the moderation more sensitive to high-difficulty instances, while still assigning nonzero update mass to hard examples.

Implementation hyperparameters. Unless otherwise specified, WELL uses $M = 10$ base learners and Hoeffding adaptive trees as the default base learner. The perturbation scale in the diversification operator is set to $\sigma = 0.05$ for all experiments. For the 19 representative synthetic imbalance settings, we use $\theta = 0.95$, $\theta_{\text{fair}} = 0.999$, $\gamma_{\text{dec}}^{\max} = 46.0$, $\alpha_{\max} = 5.0$, $k_\chi = 15.0$, and buffer size $B = 5000$. For real-world streams, we use a shorter memory and buffer, with $\theta = 0.90$, $\theta_{\text{fair}} = 0.99$, $\gamma_{\text{dec}}^{\max} = 10.0$, $\alpha_{\max} = 10.0$, $k_\chi = 15.0$, and $B = 2000$. These settings are fixed within each evaluation group and are shared by WELL and its ablated variants; each ablation only removes the corresponding weighting component.

I Additional Results for RQ2

I.1 Robustness and Additional Mechanism Checks

Sensitivity protocol and local robustness. We use a one-factor-at-a-time protocol so that each reported change has a clear parameter interpretation. For each coefficient, the local range is $\{0.6, 0.8, 1.0, 1.2, 1.4\}$; the buffer range is $B \in \{3000, 4000, 5000, 6000, 7000\}$; and the slope varied over $\{5, 10, 15, 20, 25\}$. All other quantities remain at their defaults. Each value is evaluated with ten independent runs on every setting from S0 to S18. For each stream and parameter value, we first average the ten repetitions, compute the population standard deviation across the five local values, and then average that deviation equally over the 19 streams. Table 7 reports the resulting dispersion. The largest average deviations across all five parameters are 0.0044 for BA, 0.0096 for Kappa, 0.0045 for SPD, 0.0124 for DI, and 0.0104 for GM_4. These values show local robustness around the shared defaults without implying complete parameter invariance.

Table 7: Mean within-stream standard deviation across local parameter values on S0–S18. Smaller values indicate less variation.

Parameter	Values	BA	Kappa	SPD	DI	GM_4
β	0.6 / 0.8 / 1.0 / 1.2 / 1.4	0.0021	0.0042	0.0020	0.0092	0.0077
γ_{dist}	0.6 / 0.8 / 1.0 / 1.2 / 1.4	0.0015	0.0038	0.0034	0.0094	0.0104
γ_{dec}	0.6 / 0.8 / 1.0 / 1.2 / 1.4	0.0044	0.0096	0.0045	0.0124	0.0072
B	3k / 4k / 5k / 6k / 7k	0.0007	0.0016	0.0005	0.0021	0.0015
Difficulty slope	5 / 10 / 15 / 20 / 25	0.0016	0.0043	0.0016	0.0044	0.0036

Wide-range coefficient sensitivity. We further extend each principal coefficient to $\{0, 2, 3, 4, 5\}$ while retaining the denser local scan. The value 0 removes the corresponding term, whereas values above 1 increasingly magnify its log correction before exponentiation. Table 8 reports the exact mean GM_4 values underlying Figure 8. Setting any coefficient to 0 worsens the result relative to the unit default, with the largest change occurring when $\gamma_{\mathrm{dec}} = 0$. Within 0.6–1.4, the fixed default is only 0.0006–0.0013 above the row-wise minimum, so it is close to the best local value without per-stream tuning. Larger β and γ_{dist} values cause clearer degradation, which is expected because they increasingly amplify their exponentiated corrections, while γ_{dec} remains comparatively stable over the upper range.

Table 8: Wide-range coefficient sensitivity over S0–S18. Entries are mean GM_4, and lower is better.

Varied coefficient	0	0.6	0.8	1.0 (default)	1.2	1.4	2.0	3.0	4.0	5.0
β	0.0987	0.0898	0.0891	0.0891	0.0883	0.0902	0.1011	0.1265	0.1347	0.1604
γ_{dist}	0.1083	0.0920	0.0878	0.0891	0.0915	0.0963	0.1059	0.1214	0.1381	0.1592
γ_{dec}	0.2051	0.0930	0.0900	0.0891	0.0887	0.0895	0.0946	0.0917	0.0885	0.0904

Difficulty-proxy alignment. We evaluate the recent-neighbor proxy on six positive-class safe/borderline/rare/outlier compositions. Each stationary stream contains 200,000 instances with five features, and each composition uses ten independent seeds. WELL uses a 5,000-instance FIFO buffer and $k = 100$, excludes the sensitive attribute from distances, and scores each instance before inserting it. The generator-defined categories are evaluation labels only. After excluding the first 5,000 scores for buffer warm-up, the type means follow Safe $<$ Borderline $<$ Rare $<$ Outlier across all 60 streams (Table 2). This validates sensitivity to the widely used taxonomy's difficulty spectrum, not exact recovery of discrete types or a claim that every rare instance is harmful. Figure 7 provides separate evidence that moderation improves aggregate accuracy and fairness under IDI. These complementary tests assess what the proxy measures and whether its use benefits learning.

Summarized challenges associated with IDI. The concise summary in Section 7 is an interpretation of controlled setting pairs that change IDI while keeping CI and SI fixed. For a comparison A→B, the statements in Table 9 describe the mean direction of change across the seven methods in accuracy (BA and Kappa), fairness (SPD and DI), and the joint GM_4 score. A deterioration means lower BA/Kappa or higher SPD/DI/GM_4. Rather than serving as separate tests of each challenge, these comparisons provide supporting evidence for three coupled patterns associated with IDI.

Table 9: Evidence supporting the three suggested challenges associated with IDI and their design implications.

Challenge	How IDI differs from CI/SI and supporting evidence	Design implication
Within-cell heterogeneity	CI and SI alter cell prevalence, whereas IDI alters informativeness among instances in the same cell. In S0→{S5,S6,S7}, only the dominant IDI type changes. Borderline dominance worsens mean accuracy and GM_4, while rare and outlier dominance worsen accuracy, fairness, and GM_4.	Cell weights cannot distinguish different reliability within one (Y, S) cell. Buffered neighborhood agreement supplies the required continuous instance-level moderation.
Compounded scarcity	CI and SI create cells with limited support, and IDI can reduce the reliability of the evidence that remains. Adding mixed IDI in S9→S16 and S10→S17 worsens mean accuracy and GM_4 even though the CI/SI levels are unchanged.	Cell correction restores influence to scarce cells, while difficulty moderation prevents this correction from indiscriminately amplifying unreliable arrivals. This pattern supports combining both levels.
Uneven group-error responses	IDI changes local reliability rather than group exposure alone, and different learners translate this change into different group errors. S17→S18 increases only mixed-IDI severity. Mean accuracy and GM_4 worsen, at least one fairness metric worsens, and the methods exhibit different accuracy–fairness responses.	Moderation limits harmful local updates while retaining cell-level fairness correction. It does not predetermine every metric direction, but WELL obtains the best GM_4 in S18.

The comparisons support this challenge-level interpretation rather than constituting independent tests of the three challenges. Combined with the motivating analysis and RQ1 findings, they highlight the four design-relevant aspects named in the main text: (Y, S) cell support, data-side association, prediction-side disparity, and within-cell reliability. These aspects underlie the design principles summarized in Figure 5. The first three WELL terms address cell support, data-side association, and prediction-side disparity, whereas $\chi(d_t)$ supplies the within-cell moderation that these statistics cannot express.

Fusion rule and controlled alternatives. Let $C = \exp(\beta \ell_t^{\mathrm{imb}})$, $F = \exp(\gamma_{\mathrm{dist}} \ell_t^{\mathrm{dist}})$, and $M = \exp(\gamma_{\mathrm{dec}} a_t^{\mathrm{dec}} \ell_t^{\mathrm{dec}})$ denote the three positive cell-level factors. WELL uses $CFM\chi(d_t)$. The arithmetic-mean alternative uses $(C + F + M)\chi(d_t)/3$, and the geometric-mean alternative uses $(CFM)^{1/3}\chi(d_t)$. All three alternatives use identical factor estimates, coefficient defaults, data streams, and evaluation protocols without rule-specific tuning, so the comparison changes only the fusion operation. The product retains the additive decomposition $\log(CFM) = \log C + \log F + \log M$, which keeps each correction identifiable in the relative cell influence analyzed in Appendix B.2. The arithmetic mean does not admit this additive log decomposition, while the geometric mean divides the combined log correction by three and therefore changes its effective scale. Empirically, the WELL product obtains the best mean GM_4 rank and is best in 17 of the 19 settings, as summarized in Table 10; the complete per-setting results are reported in Table 11. This controlled comparison supports the selected decomposable rule on the evaluated streams without claiming universal optimality.

Table 10: Fusion-rule comparison over S0–S18. Lower mean rank is better.

Fusion	Arithmetic mean	Geometric mean	WELL product
Mean GM_4 rank	2.37	2.47	**1.16**
Best-setting count	2/19	0/19	**17/19**

Table 11: Per-setting comparison of alternative fusion rules over ten runs. Entries are mean GM_4, lower is better, and bold values identify the best rule in each setting.

Setting	AM	GM	WELL	Setting	AM	GM	WELL
S0	0.129	0.126	**0.096**	S10	0.166	0.155	**0.148**
S1	0.141	0.135	**0.119**	S11	0.148	0.164	**0.067**
S2	**0.187**	0.198	0.215	S12	0.140	0.152	**0.024**
S3	0.123	0.119	**0.089**	S13	0.135	0.127	**0.075**
S4	0.107	0.110	**0.097**	S14	0.153	0.175	**0.078**
S5	0.131	0.134	**0.086**	S15	0.157	0.176	**0.069**
S6	**0.091**	0.105	0.095	S16	0.162	0.157	**0.054**
S7	0.092	0.106	**0.089**	S17	0.220	0.150	**0.094**
S8	0.133	0.142	**0.044**	S18	0.181	0.132	**0.057**
S9	0.138	0.135	**0.094**				

I.1.1 Computational Complexity and Runtime

Let C_{base} be the cost of one prediction and update by a base learner, M the ensemble size, B the fixed recent-buffer size, and d the feature dimension. WELL's three cell-level terms require $O(1)$ time, its exact fixed-buffer neighbor query and buffer update require $O(Bd)$ time, and ensemble prediction and updating require $O(MC_{\mathrm{base}})$ time. Its per-instance time and additional method memory are therefore

$$T_{\mathrm{WELL}} = O(MC_{\mathrm{base}} + Bd), \qquad S_{\mathrm{WELL}} = O(Bd). \tag{65}$$

Because B is fixed, both quantities are bounded independently of stream length. Table 12 compares this cost with the amortized costs of the evaluated methods. Here, W is a sliding or optimization window, R a refresh or optimization interval, P the swarm population, I the number of optimization iterations, q the number of generated samples per arrival, and k FairPCA's projection dimension. Terms divided by R amortize periodic operations. For FairPCA, B denotes the samples accumulated between projection refreshes rather than WELL's fixed buffer.

Table 12: Overall GM_4 ranks, amortized per-instance complexity, and measured learner runtime. Lower ranks and runtimes are better.

Method	Real-world rank	Synthetic rank (S0–S18)	Amortized time per instance	Runtime for 200k
FAHT	6.40	5.42	$O(C_{\mathrm{base}})$	9.42 s
CFDS	5.60	5.26	$O(C_{\mathrm{base}} + W)$	13.63 s
FOMOS	4.80	5.32	$O(C_{\mathrm{base}}[1 + P(I+1)W/R])$	1.24 min
CFSMOTE	4.40	3.74	$O((q+1)MC_{\mathrm{base}} + qBd)$	19.04 min
FABBOO	2.80	2.21	$O(MC_{\mathrm{base}} + B\log B)$	3.13 min
FairPCA	2.60	4.53	$O(C_{\mathrm{base}} + dk + Bd^2/R)$	8.03 min
WELL	**1.40**	**1.53**	$O(MC_{\mathrm{base}} + Bd)$	10.87 min

Runtime is measured on an Intel Core i5-13600KF using one numerical-backend thread. For each method, we measure the learner time required to process 200,000 instances, excluding stream generation, imports, metric computation, and file I/O. WELL is faster than the resampling-based CFSMOTE while methods with shorter runtimes obtain weaker overall GM_4 ranks on the synthetic and real-world evaluations. Together, the analytical and empirical results show that WELL attains the strongest overall effectiveness within a practical online runtime budget.

I.1.2 Per-arrival Reweighting versus Batch and Periodic Updating

We compare WELL's per-arrival reweighting and model updates with one external batch method and three controlled update-frequency variants. NUCFR (Li & Liu, 2022) learns sample weights once from an initial training and validation prefix using fairness- and utility-related influence functions, refits the model, and then freezes both the weights and model. Static estimates WELL's statistics and importance weights from the first B instances, trains once, and then freezes the weights and model. Frozen fixes WELL's empirical statistics after the first B instances but continues to compute weights and update the model for every subsequent arrival. Block uses the preceding block's statistics within

the current block and refreshes them only at block boundaries. WELL updates its empirical statistics, instance weight, and model after every arrival.

We evaluate all five methods on S0–S18 with $B \in \{5\text{k}, 10\text{k}, 20\text{k}, 50\text{k}, 100\text{k}\}$ over five independent runs. For each B, evaluation uses the same stream suffix from $B + 1$ through instance 200,000. Frozen, Block, and WELL all continue per-arrival model training, so their comparison isolates the effect of refreshing the empirical statistics rather than access to additional training instances. Table 13 reports mean GM_4 for $B = 100\text{k}$, where an em dash denotes a model that predicts only one class. Table 14 reports mean ranks for every prefix size and their aggregate over all $19 \times 5 = 95$ settings.

Table 13: Batch and update-frequency comparison at $B = 100\text{k}$. Entries are mean GM_4 over five runs, and lower is better.

Setting	Static	NUCFR	Frozen	Block	WELL	Setting	Static	NUCFR	Frozen	Block	WELL
S0	—	0.252	0.154	0.120	**0.100**	S10	0.599	0.462	0.177	0.219	**0.118**
S1	—	0.287	0.225	**0.109**	0.110	S11	—	0.517	0.445	0.314	**0.046**
S2	0.188	0.484	0.180	0.182	**0.179**	S12	—	0.469	0.184	0.200	**0.032**
S3	—	0.254	0.174	0.116	**0.084**	S13	—	0.317	0.144	0.147	**0.074**
S4	0.434	0.227	0.111	0.170	**0.078**	S14	—	0.621	0.482	0.491	**0.068**
S5	—	0.285	0.433	0.550	**0.086**	S15	—	0.629	0.756	0.769	**0.066**
S6	—	0.615	0.534	0.377	**0.100**	S16	—	0.509	0.215	0.236	**0.048**
S7	—	0.621	0.631	0.662	**0.100**	S17	—	0.599	0.345	0.324	**0.094**
S8	—	0.471	0.226	0.140	**0.045**	S18	—	0.637	0.336	0.320	**0.055**
S9	—	0.288	0.375	0.300	**0.092**						

Table 14: Mean GM_4 ranks over the 19 streams at each initial-prefix size. The overall row averages all 95 stream–prefix settings, and lower is better.

Initial prefix B	Static	NUCFR	Frozen	Block	WELL
5k	4.95	3.63	3.32	2.32	**1.00**
10k	4.84	3.63	3.47	2.26	**1.00**
20k	4.95	3.84	2.79	2.42	**1.00**
50k	4.95	3.37	3.05	2.58	**1.05**
100k	4.95	3.63	2.68	2.68	**1.05**
Overall (95 settings)	4.93	3.62	3.06	2.45	**1.02**

WELL obtains the best average rank at every initial-prefix size and an overall rank of 1.02 across the 95 settings. Its advantage over Frozen and Block supports recomputing the statistics and instance weight at every arrival, while its advantage over Static and NUCFR shows that a single batch-level weighting is less effective over the evaluated finite stream horizon.

I.2 Full Comparison on the 19 Synthetic Settings

Tables 15 and 16 report the complete comparison on the 19 representative synthetic stream settings. These settings cover balanced, single-source, and multi-source imbalance cases, including CI, SI, and the borderline, rare, and outlier components of IDI. The compared methods include fairness-aware online trees, projection-based and optimization-based methods, boosting-based methods, and resampling-based stream methods (Zhang & Ntoutsi, 2019; Lee et al., 2023; Pham et al., 2024; Iosifidis & Ntoutsi, 2020; Lammers et al., 2025; Ghosh et al., 2026). This full table complements the main-text results by showing per-setting behavior rather than only aggregate trends, and highlights that methods with strong parity control may still incur substantial accuracy loss under difficult or multi-source imbalance.

I.3 Full Comparison on Five Real-world Streams

Table 17 reports the full comparison on five real-world streams. Unlike the controlled synthetic settings, these datasets do not isolate CI, SI, and IDI by construction; therefore, they serve as an external check of whether the accuracy–fairness behavior observed in synthetic streams also appears under naturally entangled imbalance patterns. Across datasets, we report BA and Kappa for predictive performance, SPD and DI for group fairness, and GM_4 for the overall trade-off.

Table 15: Comparison on representative synthetic stream cases S0–S8 under imbalances. BA and kappa measure accuracy, SPD and DI measure fairness, and GM_4 summarizes the accuracy–fairness trade-off. ↑/↓ indicate larger/smaller is better. Parentheses report Wilcoxon test results against WELL at $p < 0.05$: $+$, $=$, and $-$ indicate significantly better, statistically tied, and significantly worse, respectively. CI/SI denote class and sensitive-group imbalance; B/R/O denote borderline, rare, and outlier IDI components. ○, ◐, and ● indicate balanced/absent, strong, and severe imbalance. Best and second-best GM_4 values are highlighted.

Dataset	CI	SI	B	R	O	Metric	FAHT	FairPCA	CFDS	FOMOS	FABBOO	CFSMOTE	WELL
S0	○	○	○	○	○	BA↑	0.9780 (+)	0.7240 (-)	0.9780 (+)	0.8385 (-)	0.9764 (+)	0.9810 (+)	0.9109
						Kappa↑	0.9574 (+)	0.4481 (-)	0.9565 (+)	0.6773 (-)	0.9513 (+)	0.9621 (+)	0.7825
						SPD↓	0.2452 (-)	0.1112 (-)	0.2405 (-)	0.1726 (-)	0.2300 (-)	0.2407 (-)	0.0467
						DI↓	0.3821 (-)	0.1843 (-)	0.3740 (-)	0.2435 (-)	0.3574 (-)	0.3779 (-)	0.0833
						GM_4 ↓	0.1121 (-)	0.2396 (-)	0.1114 (-)	0.2212 (-)	0.1129 (-)	0.1061 (-)	**0.0967**
S1	◐	○	○	○	○	BA↑	0.9743 (+)	0.6981 (-)	0.9799 (+)	0.8511 (-)	0.9795 (+)	0.9806 (+)	0.8811
						Kappa↑	0.9432 (+)	0.3967 (-)	0.9479 (+)	0.6707 (-)	0.9424 (+)	0.9490 (+)	0.7133
						SPD↓	0.2533 (-)	0.0934 (-)	0.2419 (-)	0.1797 (-)	0.2350 (-)	0.2406 (-)	0.0419
						DI↓	0.5120 (-)	0.2454 (-)	0.4880 (-)	0.3560 (-)	0.4716 (-)	0.4898 (-)	0.1332
						GM_4 ↓	0.1326 (-)	0.2573 (-)	0.1219 (=)	0.2422 (-)	0.1229 (=)	**0.1205 (=)**	0.1206
S2	●	○	○	○	○	BA↑	0.8035 (-)	0.5071 (-)	0.7916 (-)	0.8303 (-)	0.9745 (+)	0.8162 (-)	0.9345
						Kappa↑	0.9046 (+)	0.0199 (-)	0.9032 (+)	0.6324 (-)	0.8483 (+)	0.7074 (-)	0.7655
						SPD↓	0.2954 (-)	0.0033 (+)	0.2949 (-)	0.1767 (=)	0.2244 (-)	0.1474 (+)	0.1809
						DI↓	0.9798 (-)	0.3959 (+)	0.9799 (-)	0.6862 (=)	0.7247 (-)	0.8470 (-)	0.7038
						GM_4 ↓	0.2817 (-)	**0.1601 (+)**	0.2865 (-)	0.2994 (-)	0.1747 (+)	0.2908 (-)	0.2201
S3	○	◐	○	○	○	BA↑	0.9814 (+)	0.7268 (-)	0.9784 (+)	0.8373 (-)	0.8925 (-)	0.9810 (+)	0.9107
						Kappa↑	0.9636 (+)	0.4531 (-)	0.9580 (+)	0.6743 (-)	0.6935 (-)	0.9624 (+)	0.8118
						SPD↓	0.2423 (-)	0.1159 (-)	0.2420 (-)	0.1708 (-)	0.0322 (+)	0.2407 (-)	0.0388
						DI↓	0.3588 (-)	0.1839 (-)	0.3563 (-)	0.2340 (-)	0.0476 (+)	0.3581 (-)	0.0762
						GM_4 ↓	0.1036 (-)	0.2408 (-)	0.1091 (-)	0.2193 (-)	**0.0869 (=)**	0.1044 (-)	0.0873
S4	○	●	○	○	○	BA↑	0.9807 (+)	0.7169 (-)	0.9787 (+)	0.8348 (-)	0.9049 (-)	0.9807 (+)	0.9204
						Kappa↑	0.9645 (+)	0.4381 (-)	0.9595 (+)	0.6711 (-)	0.6878 (-)	0.9628 (+)	0.8901
						SPD↓	0.2429 (-)	0.1119 (-)	0.2410 (-)	0.1735 (-)	0.0759 (-)	0.2409 (-)	0.0613
						DI↓	0.3356 (-)	0.1708 (-)	0.3324 (-)	0.2217 (-)	0.1046 (+)	0.3345 (-)	0.1088
						GM_4 ↓	0.1021 (-)	0.2379 (-)	0.1061 (-)	0.2184 (-)	0.1280 (-)	0.1027 (-)	**0.0920**
S5	○	○	◐	○	○	BA↑	0.9366 (+)	0.6981 (-)	0.9365 (+)	0.8022 (-)	0.8505 (-)	0.9433 (+)	0.8920
						Kappa↑	0.8788 (+)	0.3972 (-)	0.8775 (+)	0.6052 (-)	0.6573 (-)	0.8871 (+)	0.7430
						SPD↓	0.2418 (-)	0.1035 (-)	0.2350 (-)	0.1588 (-)	0.0063 (+)	0.2232 (-)	0.0305
						DI↓	0.3647 (-)	0.1678 (-)	0.3527 (-)	0.2208 (-)	0.0093 (+)	0.3433 (-)	0.0599
						GM_4 ↓	0.1707 (-)	0.2400 (-)	0.1686 (-)	0.2329 (-)	**0.0425 (+)**	0.1583 (-)	0.0871
S6	○	○	○	◐	○	BA↑	0.6206 (-)	0.5807 (-)	0.6198 (-)	0.6109 (-)	0.5850 (-)	0.6773 (+)	0.6608
						Kappa↑	0.3746 (+)	0.1607 (-)	0.3789 (+)	0.2318 (-)	0.1606 (-)	0.3611 (+)	0.2988
						SPD↓	0.5869 (-)	0.0346 (-)	0.6147 (-)	0.1080 (-)	0.0062 (+)	0.1142 (-)	0.0110
						DI↓	0.7440 (-)	0.0779 (-)	0.7497 (-)	0.1973 (-)	0.0085 (+)	0.3340 (-)	0.0309
						GM_4 ↓	0.5733 (-)	0.1771 (-)	0.5804 (-)	0.2837 (-)	**0.0659 (+)**	0.3012 (-)	0.0958
S7	○	○	○	○	◐	BA↑	0.6214 (-)	0.5780 (-)	0.6222 (-)	0.6051 (-)	0.5802 (-)	0.6754 (+)	0.6558
						Kappa↑	0.3818 (+)	0.1555 (-)	0.3822 (+)	0.2201 (-)	0.1515 (-)	0.3565 (+)	0.2895
						SPD↓	0.6061 (-)	0.0376 (-)	0.6067 (-)	0.1020 (-)	0.0075 (+)	0.1140 (-)	0.0105
						DI↓	0.7450 (-)	0.0862 (-)	0.7444 (-)	0.1864 (-)	0.0103 (+)	0.3284 (-)	0.0271
						GM_4 ↓	0.5762 (-)	0.1859 (-)	0.5758 (-)	0.2783 (-)	**0.0731 (+)**	0.3007 (-)	0.0922
S8	○	○	◐	◐	◐	BA↑	0.7418 (+)	0.6213 (-)	0.7448 (+)	0.6735 (-)	0.7130 (-)	0.7494 (+)	0.7333
						Kappa↑	0.5046 (+)	0.2431 (-)	0.5042 (+)	0.3488 (-)	0.4004 (-)	0.5026 (+)	0.4367
						SPD↓	0.2152 (-)	0.0597 (-)	0.1941 (-)	0.1032 (-)	0.0065 (-)	0.1394 (-)	0.0033
						DI↓	0.4545 (-)	0.1215 (-)	0.4123 (-)	0.1784 (-)	0.0132 (-)	0.3208 (-)	0.0079
						GM_4 ↓	0.3393 (-)	0.2157 (-)	0.3219 (-)	0.2530 (-)	0.0628 (-)	0.2773 (-)	**0.0451**

Metric-wise accuracy–fairness improvements. The Adult comparison illustrates simultaneous gains in the constituent metrics. Relative to FABBOO, WELL raises BA from 0.7499 to 0.7696 and Kappa from 0.3868 to 0.4381, while reducing SPD from 0.0087 to 0.0037 and DI from 0.0242 to 0.0144 (Table 17). These improvements provide metric-wise evidence alongside the overall rank summary. Their extent remains dataset- and comparator-dependent, as shown by the complete results below.

I.4 Detailed Ablation on 19 Synthetic Settings

Tables 18 and 19 report component ablations across the 19 synthetic settings, covering group-imbalance correction, distributional-association correction, decision-disparity correction, the combined fairness terms, and instance-level difficulty moderation. These comparisons assess their complementary roles and how the benefits vary with the imbalance profile.

Difficulty-type dependence of moderation. Comparing WELL with WELL w/o d_t in Table 19 retains the three cell-level corrections and isolates the contribution of including difficulty moderation

Table 16: Comparison on representative synthetic stream cases S9–S18 under imbalances. BA and kappa measure accuracy, SPD and DI measure fairness, and GM_4 summarizes the accuracy–fairness trade-off. ↑/↓ indicate larger/smaller is better. Parentheses report Wilcoxon test results against WELL at $p < 0.05$: $+$, $=$, and $-$ indicate significantly better, statistically tied, and significantly worse, respectively. CI/SI denote class and sensitive-group imbalance; B/R/O denote borderline, rare, and outlier IDI components. ○, ◐, and ● indicate balanced/absent, strong, and severe imbalance. Best and second-best GM_4 values are highlighted.

Dataset	CI	SI	B	R	O	Metric	FAHT	FairPCA	CFDS	FOMOS	FABBOO	CFSMOTE	WELL
S9	◐	◐	○	○	○	BA↑	0.9784 (+)	0.6967 (-)	0.9797 (+)	0.8537 (-)	0.9786 (+)	0.9807 (+)	0.8926
						Kappa↑	0.9470 (+)	0.3939 (-)	0.9483 (+)	0.6712 (-)	0.9400 (+)	0.9492 (+)	0.7783
						SPD↓	0.2478 (-)	0.1015 (-)	0.2443 (-)	0.1818 (-)	0.2358 (-)	0.2408 (-)	0.0321
						DI↓	0.4657 (-)	0.2464 (-)	0.4593 (-)	0.3335 (-)	0.4404 (-)	0.4571 (-)	0.0948
						GM_4 ↓	0.1232 (-)	0.2636 (-)	0.1203 (-)	0.2375 (-)	0.1229 (-)	0.1184 (-)	**0.0951**
S10	●	●	○	○	○	BA↑	0.9710 (+)	0.5282 (-)	0.9717 (+)	0.8228 (-)	0.9371 (+)	0.9780 (+)	0.8611
						Kappa↑	0.8945 (+)	0.0868 (-)	0.8967 (+)	0.6028 (-)	0.5674 (-)	0.8935 (+)	0.7746
						SPD↓	0.2552 (-)	0.0119 (+)	0.2536 (-)	0.1714 (-)	0.0917 (-)	0.2394 (-)	0.0460
						DI↓	0.6568 (-)	0.4301 (-)	0.6558 (-)	0.5493 (-)	0.2320 (=)	0.6305 (-)	0.2420
						GM_4 ↓	0.1658 (-)	0.2185 (-)	0.1640 (-)	0.2906 (-)	0.1619 (-)	0.1540 (-)	**0.1404**
S11	◐	○	◐	◐	◐	BA↑	0.7350 (+)	0.5604 (-)	0.7349 (+)	0.6549 (-)	0.7182 (+)	0.7436 (+)	0.7150
						Kappa↑	0.5373 (+)	0.1419 (-)	0.5370 (+)	0.3353 (-)	0.4071 (-)	0.5214 (+)	0.4160
						SPD↓	0.2212 (-)	0.0279 (-)	0.2216 (-)	0.0941 (-)	0.0048 (-)	0.1397 (-)	0.0022
						DI↓	0.6152 (-)	0.2243 (-)	0.6157 (-)	0.3286 (-)	0.0118 (+)	0.4090 (-)	0.0250
						GM_4 ↓	0.3647 (-)	0.2223 (-)	0.3650 (-)	0.2933 (-)	0.0562 (=)	0.2936 (-)	**0.0556**
S12	○	◐	◐	◐	◐	BA↑	0.7217 (-)	0.6167 (-)	0.7320 (+)	0.6778 (-)	0.6554 (-)	0.7489 (+)	0.7296
						Kappa↑	0.4953 (+)	0.2322 (-)	0.4997 (+)	0.3585 (-)	0.2816 (-)	0.5018 (+)	0.4243
						SPD↓	0.2941 (-)	0.0558 (-)	0.2506 (-)	0.1097 (-)	0.0026 (-)	0.1409 (-)	0.0009
						DI↓	0.4900 (-)	0.1133 (-)	0.4507 (-)	0.1854 (-)	0.0040 (-)	0.3137 (-)	0.0020
						GM_4 ↓	0.3824 (-)	0.2097 (-)	0.3558 (-)	0.2573 (-)	0.0406 (-)	0.2767 (-)	**0.0224**
S13	◐	◐	◐	○	○	BA↑	0.9327 (+)	0.6495 (-)	0.9309 (+)	0.8058 (-)	0.8631 (-)	0.9409 (+)	0.8688
						Kappa↑	0.8438 (+)	0.3197 (-)	0.8430 (+)	0.5797 (-)	0.5927 (-)	0.8555 (+)	0.7121
						SPD↓	0.2470 (-)	0.0770 (-)	0.2510 (-)	0.1598 (-)	0.0048 (+)	0.2201 (-)	0.0162
						DI↓	0.4376 (-)	0.2593 (-)	0.4461 (-)	0.2954 (-)	0.0083 (+)	0.4086 (-)	0.0491
						GM_4 ↓	0.1930 (-)	0.2654 (-)	0.1961 (-)	0.2536 (-)	**0.0396 (+)**	0.1760 (-)	0.0760
S14	◐	◐	○	◐	○	BA↑	0.6683 (+)	0.5298 (-)	0.6694 (+)	0.5875 (-)	0.5824 (-)	0.6735 (+)	0.6567
						Kappa↑	0.4022 (+)	0.0728 (-)	0.4039 (+)	0.2056 (-)	0.1338 (-)	0.3797 (+)	0.3115
						SPD↓	0.1957 (-)	0.0131 (-)	0.1901 (-)	0.0714 (-)	0.0056 (-)	0.1179 (-)	0.0034
						DI↓	0.6054 (-)	0.1772 (-)	0.5991 (-)	0.3833 (-)	0.0077 (+)	0.3511 (-)	0.0299
						GM_4 ↓	0.3961 (-)	0.1796 (-)	0.3916 (-)	0.3105 (-)	**0.0633 (+)**	0.3061 (-)	0.0702
S15	◐	◐	○	○	◐	BA↑	0.6720 (+)	0.5312 (-)	0.6723 (+)	0.5855 (-)	0.5823 (-)	0.6677 (+)	0.6565
						Kappa↑	0.4057 (+)	0.0764 (-)	0.4072 (+)	0.2027 (-)	0.1349 (-)	0.3685 (+)	0.3102
						SPD↓	0.1883 (-)	0.0163 (-)	0.1897 (-)	0.0721 (-)	0.0055 (-)	0.1182 (-)	0.0028
						DI↓	0.5870 (-)	0.2036 (-)	0.5912 (-)	0.3968 (-)	0.0075 (+)	0.3492 (-)	0.0226
						GM_4 ↓	0.3876 (-)	0.1963 (-)	0.3887 (-)	0.3147 (-)	0.0628 (=)	0.3085 (-)	**0.0627**
S16	◐	◐	◐	◐	◐	BA↑	0.7419 (+)	0.5761 (-)	0.7450 (+)	0.6502 (-)	0.7074 (-)	0.7477 (+)	0.7170
						Kappa↑	0.5388 (+)	0.1743 (-)	0.5404 (+)	0.3239 (-)	0.3721 (-)	0.5295 (+)	0.4210
						SPD↓	0.2076 (-)	0.0386 (-)	0.1967 (-)	0.0920 (-)	0.0050 (-)	0.1382 (-)	0.0024
						DI↓	0.5378 (-)	0.2165 (-)	0.5113 (-)	0.3045 (-)	0.0109 (+)	0.3845 (-)	0.0155
						GM_4 ↓	0.3446 (-)	0.2347 (-)	0.3345 (-)	0.2883 (-)	0.0571 (-)	0.2861 (-)	**0.0501**
S17	●	●	◐	◐	◐	BA↑	0.7057 (=)	0.5019 (-)	0.7061 (=)	0.5531 (-)	0.6632 (-)	0.7310 (+)	0.7053
						Kappa↑	0.5009 (+)	0.0073 (-)	0.5027 (+)	0.1539 (-)	0.2269 (-)	0.4500 (+)	0.4380
						SPD↓	0.2219 (-)	0.0011 (+)	0.2219 (-)	0.0509 (-)	0.0051 (+)	0.1460 (-)	0.0063
						DI↓	0.8118 (-)	0.3041 (-)	0.8120 (-)	0.6861 (-)	0.0092 (+)	0.4856 (-)	0.0577
						GM_4 ↓	0.4087 (-)	0.1133 (-)	0.4082 (-)	0.3411 (-)	**0.0597 (+)**	0.3243 (-)	0.0890
S18	●	●	●	●	●	BA↑	0.6167 (-)	0.5066 (-)	0.6161 (-)	0.5228 (-)	0.5787 (-)	0.6505 (=)	0.6502
						Kappa↑	0.3154 (-)	0.0218 (-)	0.3105 (-)	0.0729 (-)	0.0874 (-)	0.2598 (-)	0.3284
						SPD↓	0.2000 (-)	0.0026 (-)	0.2064 (-)	0.0490 (-)	0.0048 (-)	0.1440 (-)	0.0017
						DI↓	0.9155 (-)	0.3855 (-)	0.9303 (-)	0.9168 (-)	0.0068 (+)	0.4029 (-)	0.0262
						GM_4 ↓	0.4729 (-)	0.1499 (-)	0.4796 (-)	0.3754 (-)	0.0603 (=)	0.3527 (-)	**0.0567**

in the evaluated update rule. In the rare-dominated joint setting S14, moderation raises BA from 0.6360 to 0.6567 and Kappa from 0.2486 to 0.3115, while reducing SPD from 0.0103 to 0.0034 and DI from 0.0398 to 0.0299. The outlier-dominated S15 shows the same four-metric improvement, with Kappa rising from 0.2482 to 0.3102 and SPD falling from 0.0101 to 0.0028. These settings therefore show improvements in both accuracy and fairness, not only in their composite rank.

The borderline-dominated joint setting S13 exhibits a different response. BA and Kappa remain nearly unchanged, while SPD increases from 0.0095 to 0.0162 and DI from 0.0294 to 0.0491. The aggregate benefit under IDI therefore coexists with type-dependent limitations. These results connect the difficulty-type dependence observed in the motivating analysis to the benefits and limits of WELL's instance-level component. The proxy validation in Table 2 establishes sensitivity to the taxonomy's difficulty spectrum, whereas these ablations assess the accuracy and fairness consequences of using that signal.

Table 17: Comparison on five real-world streams. We report BA and Kappa for accuracy, SPD and DI for fairness, and GM_4 for the overall accuracy–fairness trade-off. Values in parentheses are within-dataset ranks for the corresponding metric. ↑ indicates larger is better, and ↓ indicates smaller is better. The best and second-best GM_4 values are highlighted.

Dataset	Metric	FAHT	FairPCA	CFDS	FOMOS	FABBOO	CFSMOTE	WELL
Adult	BA↑	0.6699 (-)	0.5025 (-)	0.7032 (-)	0.6257 (-)	0.7499 (-)	0.7918 (+)	0.7696
	Kappa↑	0.4563 (+)	0.0075 (-)	0.4729 (+)	0.3175 (-)	0.3868 (-)	0.5279 (+)	0.4381
	SPD↓	0.1565 (-)	0.0032 (=)	0.1283 (-)	0.0714 (-)	0.0087 (-)	0.2106 (-)	0.0037
	DI↓	0.7838 (-)	0.2959 (-)	0.6965 (-)	0.6003 (-)	0.0242 (-)	0.5539 (-)	0.0144
	GM_4 ↓	0.3972 (-)	0.2129 (-)	0.3562 (-)	0.3384 (-)	0.1009 (-)	0.3382 (-)	**0.0823**
COMPAS	BA↑	0.6189 (-)	0.5225 (-)	0.6295 (=)	0.6264 (=)	0.6292 (=)	0.6520 (+)	0.6300
	Kappa↑	0.2778 (-)	0.0515 (-)	0.2879 (=)	0.2832 (=)	0.2834 (=)	0.3119 (+)	0.2831
	SPD↓	0.2622 (-)	0.0398 (-)	0.2187 (-)	0.2206 (-)	0.0179 (-)	0.0770 (-)	0.0072
	DI↓	0.6732 (-)	0.1187 (-)	0.5320 (-)	0.5307 (-)	0.0388 (-)	0.1603 (-)	0.0162
	GM_4 ↓	0.4804 (-)	0.2340 (-)	0.4296 (-)	0.4318 (-)	0.1393 (-)	0.2450 (-)	**0.1052**
KDD	BA↑	0.5875 (-)	0.5242 (-)	0.6091 (-)	0.5504 (-)	0.7554 (-)	0.8227 (+)	0.7703
	Kappa↑	0.2847 (-)	0.0905 (-)	0.3374 (+)	0.1775 (-)	0.3485 (+)	0.3539 (+)	0.2984
	SPD↓	0.0233 (-)	0.0038 (=)	0.0237 (-)	0.0103 (-)	0.0048 (=)	0.0895 (-)	0.0044
	DI↓	0.7355 (-)	0.5604 (-)	0.7285 (-)	0.7156 (-)	0.0284 (=)	0.4150 (-)	0.0276
	GM_4 ↓	0.2952 (-)	0.2436 (-)	0.2863 (-)	0.2739 (-)	0.0989 (=)	0.2684 (-)	**0.0977**
Census	BA↑	0.5402 (-)	0.4980 (-)	0.5726 (+)	0.5228 (-)	0.6116 (+)	0.5802 (+)	0.5466
	Kappa↑	0.3882 (+)	0.1724 (-)	0.4754 (+)	0.3453 (=)	0.5255 (+)	0.4707 (+)	0.3400
	SPD↓	0.0223 (-)	0.0011 (+)	0.0312 (-)	0.0084 (+)	0.0533 (-)	0.0483 (-)	0.0122
	DI↓	0.7367 (-)	0.4368 (-)	0.6544 (-)	0.6172 (-)	0.6131 (-)	0.5462 (-)	0.1756
	GM_4 ↓	0.2898 (-)	0.2148 (-)	0.2828 (-)	0.2472 (-)	0.2954 (-)	0.2944 (-)	**0.1890**
Dutch	BA↑	0.8036 (+)	0.6058 (-)	0.8035 (+)	0.7619 (-)	0.7481 (-)	0.8041 (+)	0.7786
	Kappa↑	0.6443 (+)	0.1979 (-)	0.6439 (+)	0.5191 (=)	0.4494 (-)	0.6138 (+)	0.4966
	SPD↓	0.3127 (-)	0.0124 (+)	0.3131 (-)	0.1522 (-)	0.0103 (+)	0.1928 (-)	0.0311
	DI↓	0.5000 (-)	0.0345 (+)	0.5004 (-)	0.3235 (-)	0.0164 (+)	0.3501 (-)	0.0636
	GM_4 ↓	0.3338 (-)	0.1345 (+)	0.3342 (-)	0.2847 (-)	**0.0941 (+)**	0.2777 (-)	0.1377

I.5 Detailed Ablation on Five Real-world Streams

Table 20 further evaluates the ablated variants on real-world streams.

Table 18: Comparison on representative synthetic stream cases S0–S8 under imbalances. BA and kappa measure accuracy, SPD and DI measure fairness, and GM_4 summarizes the accuracy–fairness trade-off. ↑/↓ indicate larger/smaller is better. Parentheses report Wilcoxon test results against WELL at $p < 0.05$: $+$, $=$, and $-$ indicate significantly better, statistically tied, and significantly worse, respectively. CI/SI denote class and sensitive-group imbalance; B/R/O denote borderline, rare, and outlier IDI components. ○, ◐, and ● indicate balanced/absent, strong, and severe imbalance. Best and second-best GM_4 values are highlighted.

Dataset	CI	SI	B	R	O	Metric	WELL w/o ℓ_t^{imb}	WELL w/o ℓ_t^{dist}	WELL w/o ℓ_t^{dec}	WELL w/o ℓ_t^{fair}	WELL w/o d_t	WELL
S0	○	○	○	○	○	BA↑	0.9149 (+)	0.9153 (+)	0.9725 (+)	0.9732 (+)	0.9067 (-)	0.9109
						Kappa↑	0.7920 (+)	0.7938 (+)	0.9446 (+)	0.9466 (+)	0.7696 (-)	0.7825
						SPD↓	0.0548 (-)	0.0595 (-)	0.2346 (-)	0.2370 (-)	0.0281 (+)	0.0467
						DI↓	0.1005 (-)	0.1026 (-)	0.3636 (-)	0.3673 (-)	0.0609 (+)	0.0833
						GM_4 ↓	0.1033 (-)	0.1057 (-)	0.1202 (-)	0.1193 (-)	**0.0807 (+)**	0.0967
S1	◐	○	○	○	○	BA↑	0.8784 (=)	0.8872 (=)	0.9752 (+)	0.9779 (+)	0.8650 (-)	0.8811
						Kappa↑	0.7102 (=)	0.7287 (=)	0.9292 (+)	0.9389 (+)	0.6849 (-)	0.7133
						SPD↓	0.0455 (-)	0.0507 (-)	0.2252 (-)	0.2373 (-)	0.0237 (+)	0.0419
						DI↓	0.1528 (-)	0.1535 (-)	0.4534 (-)	0.4767 (-)	0.0894 (+)	0.1332
						GM_4 ↓	0.1286 (-)	0.1279 (-)	0.1301 (-)	0.1267 (-)	**0.1000 (+)**	0.1206
S2	●	○	○	○	○	BA↑	0.9344 (=)	0.8909 (-)	0.9686 (+)	0.9802 (+)	0.9307 (=)	0.9345
						Kappa↑	0.7604 (=)	0.6759 (-)	0.8247 (+)	0.9013 (+)	0.7576 (-)	0.7655
						SPD↓	0.1855 (=)	0.1266 (+)	0.2194 (-)	0.2414 (-)	0.1802 (=)	0.1809
						DI↓	0.7003 (=)	0.6199 (+)	0.7117 (=)	0.8063 (-)	0.7091 (=)	0.7038
						GM_4 ↓	0.2223 (=)	0.2296 (=)	0.1861 (+)	**0.1584 (+)**	0.2248 (=)	0.2201
S3	○	◐	○	○	○	BA↑	0.9136 (+)	0.9163 (+)	0.9753 (+)	0.9736 (+)	0.9068 (-)	0.9107
						Kappa↑	0.8179 (+)	0.8231 (+)	0.9472 (+)	0.9485 (+)	0.8123 (=)	0.8118
						SPD↓	0.0458 (-)	0.0534 (-)	0.2319 (-)	0.2382 (-)	0.0225 (+)	0.0388
						DI↓	0.0891 (-)	0.1026 (-)	0.3421 (-)	0.3496 (-)	0.0472 (+)	0.0762
						GM_4 ↓	0.0932 (-)	0.0990 (-)	0.1146 (-)	0.1168 (-)	**0.0683 (+)**	0.0873
S4	○	●	○	○	○	BA↑	0.9188 (-)	0.9241 (+)	0.9759 (+)	0.9737 (+)	0.9248 (+)	0.9204
						Kappa↑	0.8943 (+)	0.8921 (=)	0.9488 (+)	0.9494 (+)	0.9084 (+)	0.8901
						SPD↓	0.0573 (+)	0.0731 (-)	0.2346 (-)	0.2383 (-)	0.0736 (-)	0.0613
						DI↓	0.1027 (+)	0.1272 (-)	0.3231 (-)	0.3269 (-)	0.1296 (-)	0.1088
						GM_4 ↓	**0.0887 (+)**	0.0985 (-)	0.1121 (-)	0.1144 (-)	0.0952 (-)	0.0920
S5	○	○	◐	○	○	BA↑	0.8933 (+)	0.8959 (+)	0.9394 (+)	0.9413 (+)	0.8873 (-)	0.8920
						Kappa↑	0.7464 (+)	0.7535 (+)	0.8765 (+)	0.8832 (+)	0.7299 (-)	0.7430
						SPD↓	0.0351 (-)	0.0440 (-)	0.2103 (-)	0.2216 (-)	0.0147 (+)	0.0305
						DI↓	0.0681 (-)	0.0840 (-)	0.3163 (-)	0.3310 (-)	0.0306 (+)	0.0599
						GM_4 ↓	0.0926 (-)	0.1019 (-)	0.1582 (-)	0.1590 (-)	**0.0627 (+)**	0.0871
S6	○	○	○	◐	○	BA↑	0.6597 (-)	0.6622 (=)	0.6010 (-)	0.6810 (+)	0.6562 (-)	0.6608
						Kappa↑	0.2979 (=)	0.3050 (+)	0.0226 (-)	0.3657 (+)	0.2899 (-)	0.2988
						SPD↓	0.0059 (+)	0.0037 (+)	0.6745 (-)	0.1045 (-)	0.0117 (-)	0.0110
						DI↓	0.0166 (+)	0.0107 (+)	0.7031 (-)	0.3357 (-)	0.0328 (-)	0.0309
						GM_4 ↓	0.0703 (+)	**0.0559 (+)**	0.6616 (-)	0.2937 (-)	0.0994 (-)	0.0958
S7	○	○	○	○	◐	BA↑	0.6547 (=)	0.6568 (=)	0.6016 (-)	0.6819 (+)	0.6510 (-)	0.6558
						Kappa↑	0.2887 (=)	0.2948 (+)	0.0237 (-)	0.3655 (+)	0.2800 (-)	0.2895
						SPD↓	0.0054 (+)	0.0035 (+)	0.6702 (-)	0.0998 (-)	0.0117 (-)	0.0105
						DI↓	0.0139 (+)	0.0088 (+)	0.7017 (-)	0.3258 (-)	0.0281 (=)	0.0271
						GM_4 ↓	0.0662 (+)	**0.0528 (+)**	0.6597 (-)	0.2880 (-)	0.0964 (-)	0.0922
S8	○	○	◐	◐	◐	BA↑	0.7333 (=)	0.7354 (+)	0.7271 (-)	0.7556 (+)	0.7349 (+)	0.7333
						Kappa↑	0.4379 (=)	0.4446 (+)	0.4140 (-)	0.5133 (+)	0.4384 (+)	0.4367
						SPD↓	0.0024 (+)	0.0130 (-)	0.0533 (-)	0.1369 (-)	0.0078 (-)	0.0033
						DI↓	0.0056 (+)	0.0312 (-)	0.1174 (-)	0.3173 (-)	0.0187 (-)	0.0079
						GM_4 ↓	**0.0380 (+)**	0.0890 (-)	0.1798 (-)	0.2722 (-)	0.0692 (-)	0.0451

Table 19: Comparison on representative synthetic stream cases S9–S18 under imbalances. BA and kappa measure accuracy, SPD and DI measure fairness, and GM_4 summarizes the accuracy–fairness trade-off. ↑/↓ indicate larger/smaller is better. Parentheses report Wilcoxon test results against WELL at $p < 0.05$: $+$, $=$, and $-$ indicate significantly better, statistically tied, and significantly worse, respectively. CI/SI denote class and sensitive-group imbalance; B/R/O denote borderline, rare, and outlier IDI components. ○, ◐, and ● indicate balanced/absent, strong, and severe imbalance. Best and second-best GM_4 values are highlighted.

Dataset	CI	SI	B	R	O	Metric	WELL w/o ℓ_t^{imb}	WELL w/o ℓ_t^{dist}	WELL w/o ℓ_t^{dec}	WELL w/o ℓ_t^{fair}	WELL w/o d_t	WELL
S9	◐	◐	○	○	○	BA↑	0.8721 (-)	0.8943 (=)	0.9772 (+)	0.9782 (+)	0.8682 (-)	0.8926
						Kappa↑	0.7385 (-)	0.7789 (=)	0.9326 (+)	0.9404 (+)	0.7358 (-)	0.7783
						SPD↓	0.0324 (=)	0.0419 (-)	0.2300 (-)	0.2397 (-)	0.0172 (+)	0.0321
						DI↓	0.1105 (-)	0.1221 (-)	0.4338 (-)	0.4489 (-)	0.0608 (+)	0.0948
						GM_4 ↓	0.1075 (-)	0.1078 (-)	0.1262 (-)	0.1242 (-)	**0.0797 (+)**	0.0951
S10	●	●	○	○	○	BA↑	0.8543 (-)	0.8634 (=)	0.9785 (+)	0.9830 (+)	0.8599 (=)	0.8611
						Kappa↑	0.7801 (+)	0.7854 (+)	0.8845 (+)	0.9026 (+)	0.7806 (+)	0.7746
						SPD↓	0.0388 (+)	0.0540 (-)	0.2261 (-)	0.2401 (-)	0.0463 (=)	0.0460
						DI↓	0.2169 (+)	0.2777 (-)	0.6063 (-)	0.6284 (-)	0.2447 (=)	0.2420
						GM_4 ↓	**0.1316 (+)**	0.1490 (-)	0.1525 (-)	0.1446 (-)	0.1404 (=)	0.1404
S11	◐	○	◐	◐	◐	BA↑	0.7163 (=)	0.7168 (+)	0.7382 (+)	0.7553 (+)	0.7100 (-)	0.7150
						Kappa↑	0.4260 (+)	0.4268 (+)	0.4702 (+)	0.5447 (+)	0.3946 (-)	0.4160
						SPD↓	0.0040 (-)	0.0111 (-)	0.0102 (-)	0.1375 (-)	0.0105 (-)	0.0022
						DI↓	0.0416 (-)	0.0631 (-)	0.0347 (-)	0.4112 (-)	0.0459 (-)	0.0250
						GM_4 ↓	0.0729 (-)	0.1046 (-)	0.0843 (-)	0.2861 (-)	0.0969 (-)	**0.0556**
S12	○	◐	◐	◐	◐	BA↑	0.7286 (-)	0.7303 (=)	0.7413 (+)	0.7551 (+)	0.7293 (=)	0.7296
						Kappa↑	0.4236 (=)	0.4296 (+)	0.4578 (+)	0.5099 (+)	0.4241 (=)	0.4243
						SPD↓	0.0056 (-)	0.0147 (-)	0.0173 (-)	0.1310 (-)	0.0039 (-)	0.0009
						DI↓	0.0125 (-)	0.0329 (-)	0.0443 (-)	0.2913 (-)	0.0089 (-)	0.0020
						GM_4 ↓	0.0582 (-)	0.0942 (-)	0.1031 (-)	0.2641 (-)	0.0490 (-)	**0.0224**
S13	◐	◐	◐	○	○	BA↑	0.8603 (-)	0.8691 (=)	0.9395 (+)	0.9450 (+)	0.8680 (=)	0.8688
						Kappa↑	0.6986 (-)	0.7128 (=)	0.8370 (+)	0.8533 (+)	0.7125 (=)	0.7121
						SPD↓	0.0164 (=)	0.0271 (-)	0.1881 (-)	0.2228 (-)	0.0095 (+)	0.0162
						DI↓	0.0548 (-)	0.0804 (-)	0.3499 (-)	0.3958 (-)	0.0294 (+)	0.0491
						GM_4 ↓	0.0804 (-)	0.0977 (-)	0.1683 (-)	0.1731 (-)	**0.0585 (+)**	0.0760
S14	◐	◐	○	◐	○	BA↑	0.6542 (-)	0.6586 (+)	0.6789 (+)	0.6876 (+)	0.6360 (-)	0.6567
						Kappa↑	0.3158 (+)	0.3218 (+)	0.3774 (+)	0.4249 (+)	0.2486 (-)	0.3115
						SPD↓	0.0034 (=)	0.0069 (-)	0.0244 (-)	0.0959 (-)	0.0103 (-)	0.0034
						DI↓	0.0336 (=)	0.0507 (-)	0.1096 (-)	0.3979 (-)	0.0398 (-)	0.0299
						GM_4 ↓	0.0726 (=)	0.0956 (-)	0.1533 (-)	0.2913 (-)	0.1038 (-)	**0.0702**
S15	◐	◐	○	○	◐	BA↑	0.6531 (-)	0.6576 (=)	0.6797 (+)	0.6875 (+)	0.6361 (-)	0.6565
						Kappa↑	0.3117 (=)	0.3186 (+)	0.3822 (+)	0.4249 (+)	0.2482 (-)	0.3102
						SPD↓	0.0034 (-)	0.0071 (-)	0.0291 (-)	0.0978 (-)	0.0101 (-)	0.0028
						DI↓	0.0296 (-)	0.0480 (-)	0.1308 (-)	0.3985 (-)	0.0449 (-)	0.0226
						GM_4 ↓	0.0704 (-)	0.0953 (-)	0.1673 (-)	0.2928 (-)	0.1066 (-)	**0.0627**
S16	◐	◐	◐	◐	◐	BA↑	0.7124 (-)	0.7191 (+)	0.7459 (+)	0.7574 (+)	0.7241 (+)	0.7170
						Kappa↑	0.4267 (+)	0.4326 (+)	0.4989 (+)	0.5475 (+)	0.4175 (=)	0.4210
						SPD↓	0.0069 (-)	0.0132 (-)	0.0407 (-)	0.1340 (-)	0.0064 (-)	0.0024
						DI↓	0.0463 (-)	0.0619 (-)	0.1360 (-)	0.3801 (-)	0.0227 (-)	0.0155
						GM_4 ↓	0.0861 (-)	0.1081 (-)	0.1653 (-)	0.2777 (-)	0.0702 (-)	**0.0501**
S17	●	●	◐	◐	◐	BA↑	0.6920 (-)	0.7018 (-)	0.7191 (+)	0.7472 (+)	0.7121 (+)	0.7053
						Kappa↑	0.4626 (+)	0.4513 (+)	0.4636 (+)	0.5287 (+)	0.3449 (-)	0.4380
						SPD↓	0.0165 (-)	0.0137 (-)	0.0119 (-)	0.1201 (-)	0.0054 (+)	0.0063
						DI↓	0.1603 (-)	0.1197 (-)	0.0825 (-)	0.5191 (-)	0.0732 (-)	0.0577
						GM_4 ↓	0.1462 (-)	0.1295 (-)	0.1106 (-)	0.2980 (-)	0.0940 (-)	**0.0890**
S18	●	●	●	●	●	BA↑	0.6282 (-)	0.6450 (-)	0.6228 (-)	0.6785 (+)	0.6445 (-)	0.6502
						Kappa↑	0.3315 (=)	0.3357 (+)	0.2978 (-)	0.4050 (+)	0.2143 (-)	0.3284
						SPD↓	0.0111 (-)	0.0079 (-)	0.0768 (-)	0.0904 (-)	0.0092 (-)	0.0017
						DI↓	0.1601 (-)	0.0898 (-)	0.6640 (-)	0.4920 (-)	0.0504 (-)	0.0262
						GM_4 ↓	0.1459 (-)	0.1151 (-)	0.3443 (-)	0.3073 (-)	0.1078 (-)	**0.0567**

Table 20: Comparison on five real-world streams. We report BA and Kappa for accuracy, SPD and DI for fairness, and GM_4 for the overall accuracy–fairness trade-off. Values in parentheses are within-dataset ranks for the corresponding metric. ↑ indicates larger is better, and ↓ indicates smaller is better. The best and second-best GM_4 values are highlighted.

Dataset	Metric	WELL w/o ℓ_t^{imb}	WELL w/o ℓ_t^{dist}	WELL w/o ℓ_t^{dec}	WELL w/o ℓ_t^{fair}	WELL w/o d_t	WELL
Adult	BA↑	0.7564 (-)	0.7793 (+)	0.7449 (-)	0.7468 (-)	0.7649 (-)	0.7696
	Kappa↑	0.4299 (=)	0.4870 (+)	0.3779 (-)	0.5229 (+)	0.4128 (-)	0.4381
	SPD↓	0.0095 (-)	0.0528 (-)	0.1053 (-)	0.1545 (-)	0.0038 (=)	0.0037
	DI↓	0.0407 (-)	0.1960 (-)	0.3457 (-)	0.6109 (-)	0.0129 (=)	0.0144
	GM_4 ↓	0.1094 (-)	0.1987 (-)	0.2879 (-)	0.3379 (-)	0.0823 (=)	**0.0823**
COMPAS	BA↑	0.6197 (-)	0.6241 (=)	0.5994 (-)	0.6362 (+)	0.6277 (=)	0.6300
	Kappa↑	0.2671 (-)	0.2811 (=)	0.2181 (-)	0.3014 (+)	0.2764 (=)	0.2831
	SPD↓	0.0148 (-)	0.0278 (-)	0.3470 (-)	0.1795 (-)	0.0153 (-)	0.0072
	DI↓	0.0356 (-)	0.0625 (-)	0.4689 (-)	0.3610 (-)	0.0348 (-)	0.0162
	GM_4 ↓	0.1336 (-)	0.1661 (-)	0.4853 (-)	0.3693 (-)	0.1330 (-)	**0.1052**
KDD	BA↑	0.7439 (-)	0.7349 (-)	0.7074 (-)	0.5868 (-)	0.7850 (+)	0.7703
	Kappa↑	0.3190 (+)	0.3516 (+)	0.2257 (-)	0.2894 (=)	0.2669 (-)	0.2984
	SPD↓	0.0053 (=)	0.0293 (-)	0.0595 (-)	0.0185 (-)	0.0048 (=)	0.0044
	DI↓	0.0444 (-)	0.2607 (-)	0.4343 (-)	0.7351 (-)	0.0293 (=)	0.0276
	GM_4 ↓	0.1107 (-)	0.2095 (-)	0.2925 (-)	0.2837 (-)	0.0989 (=)	**0.0977**
Census	BA↑	0.5491 (=)	0.5343 (-)	0.5362 (-)	0.5164 (-)	0.5770 (+)	0.5466
	Kappa↑	0.3515 (=)	0.3341 (=)	0.3389 (=)	0.3289 (=)	0.3743 (+)	0.3400
	SPD↓	0.0120 (=)	0.0164 (-)	0.0027 (+)	0.0101 (+)	0.0162 (-)	0.0122
	DI↓	0.1876 (=)	0.3245 (-)	0.0652 (+)	0.7142 (-)	0.1458 (+)	0.1756
	GM_4 ↓	0.1904 (=)	0.2306 (-)	**0.1312 (+)**	0.2645 (-)	0.1829 (+)	0.1890
Dutch	BA↑	0.7844 (+)	0.7926 (+)	0.6709 (-)	0.8166 (+)	0.7709 (-)	0.7786
	Kappa↑	0.5113 (+)	0.5469 (+)	0.1688 (-)	0.6331 (+)	0.4799 (-)	0.4966
	SPD↓	0.0215 (+)	0.0426 (-)	0.4536 (-)	0.1694 (-)	0.0398 (-)	0.0311
	DI↓	0.0443 (+)	0.0876 (-)	0.5928 (-)	0.3428 (-)	0.0791 (-)	0.0636
	GM_4 ↓	**0.1176 (+)**	0.1507 (-)	0.5313 (-)	0.2606 (-)	0.1540 (-)	0.1377