SAFE-STREAM: FAIRNESS-AWARE LEARNING IN STREAMING DATA WITHOUT SENSITIVE ATTRIBUTES VIA LATENT GROUP DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Fairness-aware learning in streaming environments is challenging because data distributions and model behavior may change over time, while the sensitive attributes required to measure demographic disparities may be unavailable to the learning system. This paper investigates whether latent groups discovered solely from non-sensitive observations can provide a useful operational structure for monitoring and adapting fairness in a data stream. We present *SAFE-Stream*, a fairness-aware streaming framework that maps non-sensitive features to a latent representation, maintains soft latent-group assignments, monitors Demographic Parity (DP) and Equal Opportunity (EO) over sliding windows, and combines latent-group monitoring with performance and fairness drift detection to trigger adaptive updates . Importantly, sensitive attributes are not provided to the learning, monitoring, or adaptation components; they are used only for post-hoc evaluation and a sparse audit of the discovered groups. We evaluate the framework on the UCI Adult dataset using multiple latent-group configurations, $K \in \{2, 3, 4, 5, 7, 9\}$, with five-fold cross-validation. Across configurations, SAFE-Stream achieves accuracy between 0.828 and 0.840 compared with 0.848 for an offline reference model using sensitive attributes. The framework obtains lower mean Demographic Parity gaps for both sex and race and a lower mean race-based Equal Opportunity gap than the reference, while the sex-based Equal Opportunity gap does not show the same improvement. The observed fairness differences should be interpreted together with the variation across folds. A sparse post-hoc audit further indicates a stronger association between the discovered latent groups and sex than race. These results provide an empirical investigation of latent groups as an operational structure for fairness monitoring when demographic attributes are unavailable, while also highlighting the limitations of treating latent groups as substitutes for protected-group identities.

Keywords: Fairness without sensitive attributes; Streaming machine learning; Latent group discovery; Online fairness monitoring; Fairness drift; Concept drift; Adaptive reweighting;

1 INTRODUCTION

Machine learning systems increasingly inform decisions about people, and a large body of work has developed group-fairness criteria and mitigation methods to limit outcome differences between demographic groups Dwork et al. (2012); Hardt et al. (2016); Mehrabi et al. (2021). Most of this work treats fairness as a property of a model trained once on a fixed dataset. Many deployed systems instead receive data continuously; as the distribution evolves, a model that was fair when trained may not stay fair later, so fairness becomes a quantity to track over time rather than certify once Zhang & Ntoutsi (2019); Iosifidis et al. (2019; 2021). Distributional change can affect performance and disparities differently: Lammers et al. (2026) show cumulative fairness scores over an entire stream can mask drift-induced changes in which groups are disadvantaged, motivating fairness monitoring that is local in time. Drift detectors such as ADWIN Bifet & Gavaldà (2007) monitor a scalar loss or accuracy signal but do not by themselves indicate whether a disparity between groups has changed.

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.

Monitoring and correcting group disparities requires knowledge of group membership, which is often unavailable: sensitive attributes may be legally restricted, costly, or inappropriate to collect continuously Lahoti et al. (2020); Chai et al. (2022). Removing sensitive attributes from the input does not remove disparities, since other features may act as proxies, and in a stream this is compounded because attributes may not accompany each arriving sample. Fairness-aware stream learners adapt tree ensembles, boosting, or preprocessing, but generally define discrimination with respect to a designated sensitive attribute observed in the stream Zhang & Ntoutsi (2019); Iosifidis et al. (2019; 2021); Zhang et al. (2021). Methods for fairness without demographic information, mainly studied offline, replace group labels with worst-case optimization Hashimoto et al. (2018), adversarial reweighting Lahoti et al. (2020), distillation Chai et al. (2022), related features Zhao et al. (2022), or proxy groups from clustering Yan et al. (2020); Grari et al. (2021); Chai et al. (2022) note that such approaches depend on assumptions about the relationship between surrogate groups and the underlying sensitive attributes. Other latent-space fairness methods manipulate learned representations but typically use sensitive attributes during training Chuang & Mroueh (2021); Wang et al. (2022); Jung et al. (2025).

We therefore ask whether latent groups discovered from non-sensitive representations can provide a usable structure for monitoring and adapting fairness in a stream when sensitive attributes are unobserved. We introduce *SAFE-Stream*: an MLP encoder maps each non-sensitive feature vector to a latent representation, and the Latent Group Discovery Mechanism (LGDM) maintains online-updated latent group centers and assigns each sample a soft membership vector combining a representation-distance signal with a loss-based signal. Using these memberships, an online monitor computes Demographic Parity and Equal Opportunity gaps across latent groups over a sliding window, while two ADWIN detectors track batch accuracy and the latent-group fairness gap. When a detector fires, or a gap exceeds a threshold, an adaptive update increases how many times the current batch is presented to the online logistic-regression predictor. Sensitive attributes are withheld from all learning, monitoring, and adaptation components; a small random sample is inspected for sex and race only for post-hoc analysis of the relationship between discovered and demographic groups. Discovered latent groups are not assumed to correspond directly to protected demographic groups – they are induced from non-sensitive observations and may capture only some aspects of the underlying population, so fairness measured with respect to them should not be read as a direct substitute for demographic fairness; the post-hoc audit instead characterizes the relationship between the two types of grouping.

We evaluate SAFE-Stream on the UCI Adult dataset with $K \in \{2, 3, 4, 5, 7, 9\}$ latent groups using five-fold cross-validation. Results are mixed: relative to an offline logistic regression including sex and race as inputs, SAFE-Stream has lower accuracy at every K (0.828–0.840 vs. 0.848); its mean Demographic Parity gaps for both sex and race and its mean race-based Equal Opportunity gap are lower, while its sex-based Equal Opportunity gap is higher. These fairness differences are of the same order as the variation across folds, and the evaluation includes no formal significance testing. In the audited sample, discovered latent groups associate more strongly with sex than race. These findings offer an empirical look at latent groups as an operational structure for fairness monitoring, while making explicit the limits of treating such groups as substitutes for demographic identity.

1.1 Contributions of the Work

The contributions of this work are as follows:

- We formulate a fairness-aware streaming setting in which sensitive attributes are unavailable to the learning, monitoring, and adaptation components, and investigate whether latent groups discovered from non-sensitive observations can provide an operational structure for fairness monitoring.
- We introduce *SAFE-Stream*, an online pipeline that combines latent-group discovery, sliding-window fairness monitoring, adaptive drift detection, and monitoring-driven model updates.
- We empirically evaluate the framework across multiple latent-group configurations on the UCI Adult dataset and analyze predictive performance and demographic fairness through post-hoc evaluation.

- We characterize the limitations of latent groups as surrogate structures by auditing their association with observed demographic groups and discussing the implications of representation quality and the absence of sensitive attributes.

2 Related Work

2.1 Fairness-aware machine learning

Most mitigation methods target group-fairness criteria such as Demographic Parity Dwork et al. (2012) or Equal Opportunity Hardt et al. (2016), spanning pre-, in-, and post-processing approaches Mehrabi et al. (2021). In-processing work adds fairness constraints during training, e.g., a covariance constraint between sensitive attribute and output Zafar et al. (2017), while reweighting schemes adjust sample weights via errors Chai & Wang (2022) or influence estimates Li & Liu (2022). A separate line manipulates latent representations: Fair Mixup interpolates across groups Chuang & Mroueh (2021), FAAP perturbs latents toward the decision boundary Wang et al. (2022), and ALFA perturbs latents adversarially with last-layer fine-tuning Jung et al. (2025). These methods are offline and assume observed sensitive attributes, which rules them out for a strict data stream; SAFE-Stream receives no sensitive-attribute information at runtime and uses none of these mechanisms.

2.2 Fairness without sensitive attributes

When sensitive attributes are unavailable, prior work bounds worst-case risk through distributionally robust optimization Hashimoto et al. (2018), trains a reweighting adversary over non-protected features Lahoti et al. (2020), or distills from an overfit teacher as an implicit reweighting signal Chai et al. (2022). Other approaches exploit correlated features Zhao et al. (2022), balance classes without sensitive labels Yan et al. (2020), or estimate surrogate attributes Grari et al. (2021), sometimes via clustering the latent space (see also the discussion in Chai et al., 2022). As Chai et al. (2022) note, these proxy-group methods depend on surrogate groups correlating with the true sensitive attribute. SAFE-Stream follows this surrogate-group strategy but derives groups *online*, through LGDM, a soft clustering of a latent representation fused with a loss-based signal in a single streaming pass rather than a static offline procedure; it inherits the same dependency on group quality, which we audit empirically in Section 5.3.

2.3 Fairness, drift, and online learning

FAHT extends Hoeffding trees with a fairness-aware split criterion Zhang & Ntoutsi (2019), Iosifidis et al. (2019) study fairness interventions for stream classifiers, FABBOO adapts online boosting under class imbalance while reducing discrimination Iosifidis et al. (2021), and FARF is a fair online random forest Zhang et al. (2021) — all of them measuring discrimination against a sensitive attribute observed in the stream, unlike our setting. For drift itself, we use River's ADWIN Montiel et al. (2021); Bifet & Gavaldà (2007) with default parameters over two per-batch signals, following the broader survey of Gama et al. (2014), without modifying or analyzing the detector. Lammers et al. (2026) show that cumulative fairness scores can mask drift and that window length and overlap affect windowed scores materially. E-FairDrift (Dwivedi & Kurmi, 2026, workshop paper; abstract only) discovers candidate subgroups from black-box signals and applies sequential tests with statistical guarantees for sparse audit labels; SAFE-Stream instead couples grouped monitoring directly to an online update rule for the predictor and offers no such guarantees, so we do not compare against it directly.

3 Problem Setup

3.1 Problem setup

We consider a streaming setting in which data arrives in mini-batches and must be processed without access to sensitive attributes at prediction time (notation summarized in Table 1). A stream is $\mathcal{D} = \{(x_t, y_t)\}_{t=1}^{T}$, $T = 30{,}139$, processed in mini-batches $B_1, \ldots, B_M$ in row order. At batch m the learner predicts $\hat{y}_t$ for $t \in B_m$, then observes y_t and updates (test-then-train, batch granularity).

Table 1: Notation used throughout the paper.

Symbol	Meaning
$x_t \in \mathbb{R}^d$, $d = 97$	non-sensitive feature vector after preprocessing
$y_t \in \{0, 1\}$	label (income >50K)
$a_t = (a_t^{\text{sex}}, a_t^{\text{race}})$	sensitive attributes, $a_t^{\text{sex}} \in \{0, 1\}$, $a_t^{\text{race}} \in \{0, \ldots, 4\}$ (withheld from predictor, encoder, and online control logic; seen only by the sparse audit buffer)
B_m, $\lvert B_m \rvert = 256$	m-th mini-batch, except possibly the last
$z_t = f_\theta(x_t) \in \mathbb{R}^{16}$	latent representation
K	number of latent groups
$c_t \in \Delta^{K-1}$	soft latent-group membership vector
$W = 500$	sliding-window length (samples)
$\varepsilon = 0.05$	fairness-violation threshold

3.2 Information available at each stage

We separate four information regimes, to avoid the ambiguity common in fairness-without-demographics work. **Learning:** (x_t, y_t) per sample, after prediction; a_t is never given to the predictor or encoder. **Monitoring:** $\hat{y}_t$, y_t, and soft latent-group memberships c_t from x_t only — no sensitive attribute involved. **Sparse audit:** for a random subset $\mathcal{A}$, $|\mathcal{A}| \leq \lfloor \rho T \rfloor$, $\rho = 0.02$, sampled without replacement from each batch, $(a_t^{\text{sex}}, a_t^{\text{race}})$ is buffered with the sample's latent group — the only point where a_t is observed. The buffer is read once, after the stream ends, to estimate each latent group's sensitive-attribute composition (Section 4.7); it affects no prediction, drift call, or reweighting decision already made online. **Post-hoc evaluation:** a_t for every held-out sample in cross-validation, used only to compute the reported fairness metrics in Section 5.2; the 2% audit is a separate, diagnostic resource.

3.3 Objective and fairness quantities

We want a predictor with high accuracy whose disparities across the true groups $\{a_t = g\}$ stay small over the stream, without observing a_t. SAFE-Stream instead monitors disparities across K latent groups derived from x_t and adapts when they grow; whether these track disparities across true groups is examined empirically in Section 5.3, with no guarantee that they do.

For a partition into groups $\mathcal{G}$ (true or latent) and predictions $\hat{y}$, define

$$s_g = \mathbb{P}(\hat{y} = 1 \mid g), \qquad \tau_g = \mathbb{P}(\hat{y} = 1 \mid g,\, y = 1), \tag{1}$$

giving the Demographic Parity and Equal Opportunity gaps

$$\Delta_{\text{DP}} = \max_{g \in \mathcal{G}} s_g - \min_{g \in \mathcal{G}} s_g, \qquad \Delta_{\text{EO}} = \max_{g \in \mathcal{G}} \tau_g - \min_{g \in \mathcal{G}} \tau_g, \tag{2}$$

with groups lacking positive samples skipped in Δ_{EO}; these reduce to absolute differences for two groups and to max–min ranges otherwise. Here Δ_{EO} denotes Equal Opportunity, not Equalized Odds, which would also include false-positive rates.

4 SAFE-Stream

4.1 Overview and streaming setting

Figure 1 summarizes the setup. The UCI Adult dataset is cleaned and encoded into features X and label y, with sex and race retained only as sensitive attributes for fairness evaluation. Two branches are evaluated on the same cross-validation folds: *SAFE-Stream*, which withholds sensitive attributes from the model, and the *true-baseline*, an offline model that uses them explicitly as input; both are scored in one comparison table (performance and fairness).

Each SAFE-Stream batch is processed by Algorithm 1: (1) embed the batch, (2) compute soft latent-group memberships with LGDM, (3) predict with the online predictor, (4) update group loss statistics, (5) append predictions/labels/memberships to a sliding window and compute latent-group gaps,

(6) update two ADWIN detectors, (7) check the trigger and, if fired, apply adaptive reweighting, (8) draw the sparse audit sample, and (9) perform the regular predictor update.

The stream is cut into batches of 256 samples in row order (118 batches full stream; ∼95 per CV fold). Predictions within a batch use parameters from the previous batch; after monitoring/adaptation, the predictor takes one regular update per sample (update interval 1).

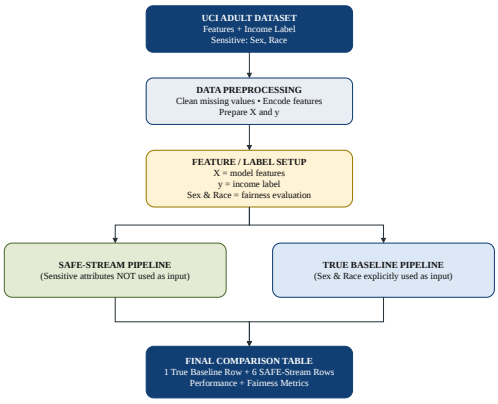


Figure 1: SAFE-Stream pipeline: the UCI Adult dataset is preprocessed and then split into a SAFE-Stream branch (sensitive attributes withheld from the model) and a true-baseline branch (sensitive attributes used explicitly as input); both are evaluated in a single comparison table.

4.2 Latent Representation

$$z = f_\theta(x) = W_2\,\mathrm{ReLU}(W_1 x + b_1) + b_2, \qquad W_1 \in \mathbb{R}^{32\times 97},\ W_2 \in \mathbb{R}^{16\times 32},$$

is a fixed random projection: θ is randomly initialized (PyTorch default, seed 42) and never updated, so z only defines the geometry used for latent-group discovery. The online predictor operates on x directly (Sec. 4.6), not on z.

4.3 Latent Group Discovery Mechanism (LGDM)

LGDM keeps centers $\mu_1, \ldots, \mu_K \in \mathbb{R}^{16}$ and running group losses $L_1, \ldots, L_K$ (init. 0). For batch B with embeddings $\{z_i\}$:

Init (first batch only). $\mu_k = z_{i_k}$, $i_k = \lfloor (k-1)(|B|-1)/(K-1) \rfloor$.

Center update. $h_i = \arg\min_k \|z_i - \mu_k\|_2$; for $S_k = \{i : h_i = k\} \neq \emptyset$,

$$\mu_k \leftarrow (1-\eta_\mu)\mu_k + \eta_\mu \frac{1}{|S_k|}\sum_{i\in S_k} z_i, \qquad \eta_\mu = 0.2.$$

Representation signal. With $d_{ik} = \max(\|z_i - \mu_k\|_2, 10^{-8})$,

$$r_{ik} = \frac{1/d_{ik}}{\sum_j 1/d_{ij}}.$$

Loss-based signal. $\ell_k = 1/K$ on the first batch, else

$$\ell_k = \frac{1 + \max(L_k, 0)}{\sum_j (1 + \max(L_j, 0))};$$

after scoring, $L_k \leftarrow 0.9 L_k + 0.1\bar{\ell}_m$ for all k, where $\bar{\ell}_m$ is the mean batch log-loss.

Fusion. $c_i = 0.6\, r_i + 0.4\, \ell \in \Delta^{K-1}$; $\arg\max c_i$ labels audit records only (Sec. 4.7).

Observation 1 (loss signal is group-uniform)**.** *$L_1 = \cdots = L_K$ at all times, since every L_k starts at 0 and receives the same affine update from the same $\bar{\ell}_m$. Hence $\ell_k = 1/K$ and $c_i = 0.6\, r_i + (0.4/K)\mathbf{1}$: all group structure in our results derives from the representation signal on the fixed embedding. A group-specific loss signal, e.g. $L_k \propto \sum_i c_{ik}\ell_i$, is left to future work.*

Observation 2 (shrinkage)**.** *Since $c_{ik} \geq 0.4/K$, any group-conditional statistic weighted by c_{ik} (Sec. 4.4) is a convex combination of the corresponding r_{ik}-weighted statistic and the window-wide mean. Latent-group gaps are therefore attenuated relative to gaps computed from r alone, for any weights of the form $0.6\, r + 0.4/K$, independent of the data.*

Centers move every batch ($\eta_\mu = 0.2$); K is fixed. There is no mechanism for matching group identities across time or for creating/merging/deleting groups, so "group k" at different times need not refer to the same latent-space region. Per-group counts are tracked but unused by the assignment.

4.4 Online fairness monitoring and drift detection

A FIFO window $\mathcal{W}$ holds the most recent $W = 500$ triples $(c_i, \hat{y}_i, y_i)$ ($\sim$two batches; overlapping). Once full,

$$\hat{s}_k = \frac{\sum_{i\in\mathcal{W}} c_{ik}\hat{y}_i}{\sum_{i\in\mathcal{W}} c_{ik}}, \qquad \hat{\tau}_k = \frac{\sum_{i\in\mathcal{W}} c_{ik}\,\mathbb{I}[y_i = 1]\,\hat{y}_i}{\sum_{i\in\mathcal{W}} c_{ik}\,\mathbb{I}[y_i = 1]},$$
$$\hat{\Delta}_{\mathrm{DP}} = \max_k \hat{s}_k - \min_k \hat{s}_k, \qquad \hat{\Delta}_{\mathrm{EO}} = \max_k \hat{\tau}_k - \min_k \hat{\tau}_k,$$

the soft-membership analogues of Sec. 3.3. The window includes current-batch predictions made before the update; both gaps are undefined until it is full.

Two instances of River's ADWIN Montiel et al. (2021) (default parameters) run once per batch: over batch accuracy $a_m = \frac{1}{|B_m|}\sum_{t\in B_m}\mathbb{I}[\hat{y}_t = y_t]$ (performance), and over $g_m = \max\{\hat{\Delta}_{\mathrm{DP}}, \hat{\Delta}_{\mathrm{EO}}\}$ once defined (fairness). A violation is flagged when $\hat{\Delta}_{\mathrm{DP}} > \varepsilon$ or $\hat{\Delta}_{\mathrm{EO}} > \varepsilon$, $\varepsilon = 0.05$; the adaptation trigger is

$$\mathrm{trigger}_m = \mathrm{drift}_m^{\mathrm{perf}} \ \vee\ \mathrm{drift}_m^{\mathrm{fair}} \ \vee\ \mathrm{violation}_m,$$

so adaptation also fires on a threshold, independently of ADWIN. Per-trigger-type counters are recorded but not printed by the notebook, so their relative frequency cannot be reported.

4.5 Adaptive update intensification

$\lambda \in \mathbb{R}^K_{\geq 0}$, initialized to 0. On a triggered batch with $v = \max\{\hat{\Delta}_{\mathrm{DP}}, \hat{\Delta}_{\mathrm{EO}}\}$ over the defined values (step skipped if neither is defined),

$$\lambda \leftarrow \max\Big(\lambda + \eta\, v \sum_{i\in B} c_i,\ 0\Big),\ \eta = 0.01; \qquad w_i = 1 + c_i^\top \lambda;$$
$$\omega = 1 + \frac{1}{|B|}\sum_{i\in B} w_i, \qquad r = \min\big(\max(1, \lceil\omega\rceil), 10\big).$$

Each sample of the batch is then presented to the predictor r consecutive times (sample by sample, in batch order), rather than as r separate passes over the batch. λ is non-decreasing over the stream; there is no decay or projection step.

Observation 3. *The vector λ is updated before the weights are computed. Since $c_{ik} \geq 0.4/K > 0$, whenever $v > 0$ every $\lambda_k > 0$, so $w_i > 1$, $\omega > 2$ and $r \geq 3$. With the regular update, a triggered batch is therefore presented to the predictor 4–11 times. Since r depends only on the batch mean of w_i, all samples of a batch receive the same weight: reweighting does not differentiate between samples or groups within a batch, and is better described as an adaptive intensification of the update than as per-sample reweighting.*

4.6 Online model adaptation

The online predictor is River's `StandardScaler | LogisticRegression`, SGD at learning rate 0.005 (other options at River defaults), on the 97-dimensional preprocessed features x_t; $\hat{y}_t = \mathbb{I}[p(y = 1 \mid x_t) \geq 0.5]$. A regular update applies one `learn_one` step to every sample of every batch; triggered batches receive r additional steps per sample (Sec. 4.5).

4.7 SPARSE AUDIT AND LATENT-GROUP CHARACTERIZATION

Per batch, the audit draws $\min(\text{remaining}, \max(1, \lceil 0.02|B| \rceil))$ samples uniformly without replacement (6 for a full batch), until a budget of $\lfloor 0.02\, N_{\text{stream}} \rfloor$ is reached (602 samples, exhausted at batch 101/118, so the audit covers only the earlier part of the stream). Each audited sample stores its argmax latent group, membership vector, label, sex and race codes. After the stream, the notebook computes, per latent group, the sex and race proportions among audited samples and an "anchor strength" (largest single proportion). None of this feeds back into assignment, monitoring, drift detection, reweighting or prediction — it is post-hoc validation only (Sec. 5.3). Because anchor strength is a maximum proportion, for race it approximately tracks the majority-category (White) share and so reflects base rates as much as association; we therefore report proportions directly.

Algorithm 1 SAFE-Stream (as implemented)

Input: stream batches $B_1, \ldots, B_M$; K; W=500; ε=0.05; η=0.01; audit fraction ρ=0.02
State: fixed encoder f_θ; centers $\mu_{1:K}$; group losses L_k=0; λ=0; FIFO window $\mathcal{W}$ (size W); $\text{ADWIN}_{\text{perf}}$, $\text{ADWIN}_{\text{fair}}$; predictor g
for $m = 1, \ldots, M$ **do**
  $z_i \leftarrow f_\theta(x_i)$; $c_i \leftarrow \text{LGDM}(z_i; \mu, L)$ (centers updated inside); $\hat{y}_i \leftarrow g(x_i)$ ▷ predictions before learning
  $a_m \leftarrow$ batch accuracy; $\bar{\ell}_m \leftarrow$ mean log-loss; $L_k \leftarrow 0.9 L_k + 0.1 \bar{\ell}_m \; \forall k$
  append $(c_i, \hat{y}_i, y_i)$ to $\mathcal{W}$; if $|\mathcal{W}|$=W, compute $\hat{\Delta}_{\text{DP}}, \hat{\Delta}_{\text{EO}}$
  update $\text{ADWIN}_{\text{perf}}$ with a_m; update $\text{ADWIN}_{\text{fair}}$ with $\max(\hat{\Delta}_{\text{DP}}, \hat{\Delta}_{\text{EO}})$ if defined
  trigger $\leftarrow \text{drift}_{\text{perf}} \vee \text{drift}_{\text{fair}} \vee (\hat{\Delta}_{\text{DP}} > \varepsilon \vee \hat{\Delta}_{\text{EO}} > \varepsilon)$
  if trigger **and** a gap is defined **then**
    $\lambda \leftarrow \lambda + \eta \cdot v \cdot \sum_i c_i$; $\omega \leftarrow 1 + \text{mean}_i(1 + c_i^\top \lambda)$; $r \leftarrow \min(\max(1, \lceil \omega \rceil), 10)$
    for each $i \in B_m$: repeat r times g.`learn_one`(x_i, y_i)
  end if
  draw sparse audit sample (records latent group, sex, race; not used above)
  g.`learn_one`(x_i, y_i) for all $i \in B_m$ ▷ regular update
end for

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Dataset and preprocessing. We use the UCI Adult dataset Becker & Kohavi (1996) (`adult.data`; 32,561 rows). Rows with missing values are dropped (30,162 remain) and 23 duplicates removed, leaving 30,139 samples. The target is income >50K (1) vs. ≤50K (0). Sensitive attributes are sex (Female 0, Male 1) and race (White 0, Black 1, Asian-Pac-Islander 2, Amer-Indian-Eskimo 3, Other 4); both are excluded from the model input. The 12 non-sensitive features (`age`, `workclass`, `fnlwgt`, `education`, `education-num`, `marital-status`, `occupation`, `relationship`, `capital-gain`, `capital-loss`, `hours-per-week`, `native-country`) are standardized (numeric) / one-hot encoded (categorical), giving $d = 97$. The preprocessor is fit on the training fold only (Protocol A) or on the whole dataset (Protocol B). Excluding sex and race does not remove proxies: `relationship`/`marital-status` likely correlate with sex and `native-country` with race . **Streaming protocols.** *Protocol A (5-fold CV):* stratified 5-fold split (shuffled, seed 42) on y; for each fold and K, a fresh seed-42 encoder processes the training fold in row order, and the final predictor is evaluated on the held-out fold with true sex/race labels. *Protocol B (full stream):* for each K, all 30,139 samples are processed once in row order with a fresh seed-42 encoder; we record prequential accuracy/log-loss, end-of-stream latent-group gaps, and a sparse audit (602 samples). The stream order is the cleaned row order of `adult.data`; there is no shuffling within a training stream, no injected drift, and no timestamp in Adult, so any drift present is incidental and untested.

Configuration. Latent groups $K \in \{2, 3, 4, 5, 7, 9\}$; encoder: Linear(97,32)–ReLU–Linear(32,16), random init (seed 42), untrained; LGDM center rate 0.2, fusion 0.6/0.4 (rep./loss), loss EMA 0.9/0.1; batch size 256; window W=500, ε=0.05; ADWIN (River defaults, two instances); adaptive update intensification η=0.01, repetition cap 10; predictor: River

Table 2: 5-fold CV, mean ± s.d. over folds. Lower is better for DP/EO.

K	Acc.	F1	ROC-AUC	Sex DP	Sex EO	Race DP	Race EO
–	0.8479±0.0067	0.6649±0.0108	0.9045±0.0067	0.1832±0.0047	0.0955±0.0110	0.2175±0.0578	0.5033±0.1820
2	0.8399±0.0076	0.6380±0.0134	0.8939±0.0065	0.1803±0.0110	0.1326±0.0372	0.2048±0.0609	0.4183±0.1786
3	0.8401±0.0069	0.6385±0.0115	0.8941±0.0062	0.1812±0.0117	0.1335±0.0371	0.2071±0.0647	0.4183±0.1786
4	0.8399±0.0080	0.6379±0.0149	0.8931±0.0077	0.1813±0.0085	0.1388±0.0341	0.2005±0.0700	0.4383±0.1694
5	0.8346±0.0075	0.6243±0.0183	0.8875±0.0085	0.1725±0.0196	0.1197±0.0338	0.2089±0.0799	0.4436±0.1606
7	0.8281±0.0078	0.6107±0.0125	0.8791±0.0092	0.1695±0.0200	0.1060±0.0376	0.2050±0.0894	0.4280±0.1437
9	0.8317±0.0087	0.6184±0.0195	0.8838±0.0085	0.1709±0.0220	0.1098±0.0445	0.1897±0.0487	0.4327±0.1262

Table 3: Full stream (30,139 samples, 118 batches), one run per K. Last two columns are latent-group (not demographic) gaps on the final window.

K	Preq. Acc.	Log-loss	Final $\hat{\Delta}_{\mathrm{DP}}$ (latent)	Final $\hat{\Delta}_{\mathrm{EO}}$ (latent)	Audited
2	0.8365	0.3557	0.0268	0.0100	602
3	0.8365	0.3557	0.0323	0.0208	602
4	0.8338	0.3623	0.0326	0.0356	602
5	0.8353	0.3608	0.0488	0.0656	602
7	0.8312	0.3699	0.0356	0.0338	602
9	0.8294	0.3767	0.0508	0.0682	602

`StandardScaler|LogisticRegression` (SGD, lr 0.005, threshold 0.5); update interval 1 batch; audit fraction 2% (random, unseeded, per-batch).

Metrics. Predictive: accuracy, F1, ROC-AUC per fold (A); prequential accuracy and mean log-loss (B).Fairness: sex and race evaluated separately via Δ_{DP}, Δ_{EO} (Section 3.3), mean ± sample s.d. over 5 folds (A); final-batch latent-group gaps $\hat{\Delta}_{\mathrm{DP}}$, $\hat{\Delta}_{\mathrm{EO}}$ (B), not comparable in scale to demographic gaps.

Baseline. An offline scikit-learn logistic regression (max. 1,000 iters, seed 42) trained in batch mode on the same five training folds, *including* sex and race as inputs. It is not a fairness intervention; it is a reference for accuracy and for the disparities of an unconstrained model with access to sensitive attributes. No online, non-fairness, or fairness-without-demographics baseline is included.

5.2 Results

5.2.1 Cross-validation (Protocol A)

SAFE-Stream's accuracy (0.828–0.840), F1 (0.611–0.639) and ROC-AUC (0.879–0.894) trail the baseline (0.848/0.665/0.905) in every fold and K (gap 0.006–0.027). Performance is similar for $K \in \{2, 3, 4\}$ (≈0.840) and lower for $K \in \{5, 7, 9\}$ (0.835/0.828/0.832), non-monotone in K. Mean sex and race DP gaps are lower than the baseline at every K (sex: by 0.002–0.014; race: 0.1897–0.2089 vs. 0.2175), but fold-level s.d. (sex 0.009–0.022; race 0.049–0.089) is comparable to or larger than these differences, and no significance test was applied. Sex EO is *higher* than baseline at every K (0.106–0.139 vs. 0.0955); race EO is lower at every K (0.418–0.444 vs. 0.5033) but highly variable across folds (a max–min range over 5 categories, some very small). Thus sex DP improves slightly while sex EO worsens; race DP and EO both decrease in the mean but with far larger fold-to-fold variation. K=2 and K=3 give identical fold-level results in folds 2 and 5, and $K \in \{2, 3, 4\}$ coincide in fold 3, indicating the same final predictor in those cases (trigger counts were not logged, so the cause is unknown).Per-fold values for accuracy, F1, ROC-AUC, sex DP, and race DP are reported in Appendix A.1 (Table 4).

5.2.2 Full-stream run (Protocol B)

Prequential accuracy falls from 0.8365 (K=2, 3) to 0.8294 (K=9) and log-loss rises from 0.3557 to 0.3767 (Table 3). Final latent-group gaps (0.010–0.068) are much smaller than the demographic gaps in Table 2 (≈0.17–0.18 for sex DP); this may partly reflect representation attenuation, but latent

and demographic groups are different partitions and both are single end-of-stream figures, so they are not directly comparable.

5.3 Analysis: What Do the Latent Groups Capture?

The 602 audited samples (Protocol B) give the only direct evidence linking latent groups to sex/race (per-K, per-group counts and proportions in Appendix A.2). Sex composition varies markedly across groups (female share 0.074–0.577 vs. an overall audited share of $\sim$0.30–0.36; e.g., at K=3 one group is 92.6% male, another only 57.0% male), with Cramér's V of 0.21–0.30 across K — a modest but non-negligible association. Race composition stays closer to the overall sample (White share 0.80–0.94 in 26/30 groups, vs. an overall $\sim$0.85–0.90; four exceptions), with Cramér's V of 0.05–0.14; non-White/Black categories have very small counts, so group-level race proportions are imprecise.Cramér's V and the overall audited shares are computed from the audited per-group counts (Appendix A.2 lists the Female, White and Black shares).

Latent groups from a fixed random embedding of non-sensitive features are thus moderately informative about sex and much less informative about race, consistent with concerns raised for surrogate-group methods Chai et al. (2022). Since the audit is a single, unseeded sample per K with small group sizes at larger K, and latent-group gap trajectories were not compared to demographic gap trajectories, this evidence is descriptive: we do not show that latent-group drift tracks demographic drift.

Relation to Table 2. SAFE-Stream differs from the baseline in three ways at once (excludes sensitive attributes, learns online in one pass via SGD, applies monitoring-driven adaptation), so the results do not isolate the individual contributions of LGDM, monitoring, or reweighting.

Limitations. The encoder is fixed/untrained and the predictor ignores the latent representation, so LGDM groups form mainly from the encoder signal. The 2% audit is descriptive only and reweighting is batch-level, not sample-level; ADWIN trigger frequency and group identities over time were not tracked. Adult has no injected drift and fairness-over-time is not reported prequentially. The only baseline is an offline sex/race model, with no ablations or online fairness comparison; each configuration ran once per fold with no significance testing, and race gaps are sensitive to small subgroups. Features may encode sex/race proxies, and reproducibility is limited by unseeded audit sampling and an unpinned River version. Overall, component contributions are not isolated, and latent-group gaps should not be treated as demographic-fairness proxies — especially for race.

6 Conclusion

This work asked whether groups discovered from non-sensitive observations can provide a useful operational structure for monitoring and adapting fairness in streaming machine learning when sensitive attributes are unavailable to the learner. We introduced *SAFE-Stream*, combining latent-group discovery, sliding-window fairness monitoring, adaptive drift detection, and monitoring-driven updates without giving demographic attributes to the learning or adaptation process. On the UCI Adult dataset, SAFE-Stream achieved accuracy between 0.828 and 0.840 against 0.848 for an offline reference model using sex and race, with lower mean Demographic Parity gaps for both sex and race and a lower mean race-based Equal Opportunity gap, though the sex-based Equal Opportunity gap did not improve; these differences should be read alongside the cross-fold variation, since no formal significance testing was performed. The post-hoc audit showed the latent groups associate more strongly with sex than with race, so the representation captures some demographic structure but is not a reliable stand-in for protected-group identity – latent groups can support operational monitoring, but fairness with respect to them should not be equated with demographic fairness. Overall, this is an initial assessment of latent-group discovery for fairness monitoring without demographic information, constrained by the fixed representation, the simplified loss-based group signal, batch-level update intensification, and the descriptive audit.

Future work. Priorities include training the encoder (or comparing embeddings) and adding a group-specific loss signal; using the audited sample to calibrate or map latent groups to demographic groups, with a decaying reweighting multiplier; and prequential, windowed evaluation on streams with injected drift and multiple stream orders, alongside correlation analysis between latent-group and demographic gaps over time.

AI Use Statement: In this work, we used generative AI tools solely to improve the readability, grammar, and clarity of the manuscript. We reviewed and verified all AI-assisted text. The research ideas, conceptual framework, methodology, experimental design, implementation, analysis, results, and conclusions were developed and verified by the authors. We take full responsibility for the final content of this work.

REFERENCES

B. Becker and R. Kohavi. Adult. UCI Machine Learning Repository, 1996.

A. Bifet and R. Gavaldà. Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pp. 443–448, 2007. doi: 10.1137/1.9781611972771.42.

J. Chai and X. Wang. Fairness with adaptive weights. In *International Conference on Machine Learning*, pp. 2853–2866, 2022.

J. Chai, T. Jang, and X. Wang. Fairness without demographics through knowledge distillation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19152–19164, 2022.

C.-Y. Chuang and Y. Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021.

R. R. Dwivedi and V. K. Kurmi. E-FairDrift: Anytime-valid tests for intersectional fairness drift. Poster at the ICML 2026 Hypothesis Testing Workshop, 2026.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44, 2014. doi: 10.1145/2523813.

V. Grari, S. Lamprier, and M. Detyniecki. Fairness without the sensitive attribute via causal variational autoencoder. arXiv:2109.04999, 2021.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Tatsunori Hashimoto, Meghana Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1929–1938, 2018.

V. Iosifidis, T. N. H. Tran, and E. Ntoutsi. Fairness-enhancing interventions in stream classification. In *Database and Expert Systems Applications (DEXA)*, pp. 261–276, 2019. doi: 10.1007/978-3-030-27615-7_20.

V. Iosifidis, W. Zhang, and E. Ntoutsi. Online fairness-aware learning with imbalanced data streams. arXiv:2108.06231, 2021.

H. Jung, J. Chai, and X. Wang. Adversarial latent feature augmentation for fairness. In *International Conference on Learning Representations*, 2025.

P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi. Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Kathrin Lammers, Fabian Hinder, Barbara Hammer, and Valerie Vaquet. Drift-aware evaluation of fair stream learning. In *Proceedings of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pp. 127–132, 2026. doi: 10.14428/esann/2026.ES2026-152.

P. Li and H. Liu. Achieving fairness at no utility cost via data reweighing with influence. In *International Conference on Machine Learning*, pp. 12917–12930, 2022.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021. doi: 10.1145/3457607.

Jacob Montiel, Max Halford, Saulo Martiello Mastelini, Geoffrey Bolmier, Raphael Sourty, Robin Vaysse, Adil Zouitine, Heitor Murilo Gomes, Jesse Read, Talel Abdessalem, and Albert Bifet. River: Machine learning for streaming data in Python. *Journal of Machine Learning Research*, 22(110):1–8, 2021.

Z. Wang, X. Dong, H. Xue, Z. Zhang, W. Chiu, T. Wei, and K. Ren. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10379–10388, 2022.

S. Yan, H.-T. Kao, and E. Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 1715–1724, 2020. doi: 10.1145/3340531.3411980.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970, 2017.

W. Zhang and E. Ntoutsi. FAHT: An adaptive fairness-aware decision tree classifier. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1480–1486, 2019. doi: 10.24963/ijcai.2019/205.

Wenbin Zhang, Albert Bifet, Xiangliang Zhang, Jeremy C. Weiss, and Wolfgang Nejdl. FARF: A fair and adaptive random forests classifier. In *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021, Proceedings*, volume 12713 of *Lecture Notes in Computer Science*, pp. 245–256. Springer, 2021. doi: 10.1007/978-3-030-75765-6_20.

Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 1433–1442. Association for Computing Machinery, 2022. doi: 10.1145/3488560.3498493.

A Appendix

A.1 Per-fold results (Protocol A)

Baseline (per fold). Accuracy: 0.8499, 0.8553, 0.8504, 0.8374, 0.8464. F1: 0.6676, 0.6778, 0.6664, 0.6477, 0.6647. ROC-AUC: 0.9033, 0.9123, 0.9075, 0.8941, 0.9052. Per-fold baseline fairness values were computed by the notebook but not printed The baseline means $\pm$ standard deviations are in Table 2.

Table 4: Per-fold results of SAFE-Stream (Protocol A), folds 1–5.

Metric	K	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	2	.8442	.8456	.8436	.8270	.8394
	3	.8434	.8456	.8436	.8285	.8394
	4	.8437	.8461	.8436	.8261	.8399
	5	.8328	.8414	.8427	.8247	.8314
	7	.8296	.8280	.8378	.8160	.8291
	9	.8431	.8296	.8364	.8198	.8296
F1	2	.6495	.6451	.6472	.6184	.6300
	3	.6478	.6451	.6472	.6226	.6300
	4	.6480	.6488	.6472	.6147	.6310
	5	.6258	.6376	.6460	.6101	.6022
	7	.6242	.6053	.6241	.6012	.5986
	9	.6481	.6058	.6282	.6082	.6018
ROC-AUC	2	.8932	.9004	.8990	.8839	.8932
	3	.8929	.9004	.8990	.8847	.8932
	4	.8928	.8999	.8990	.8806	.8932
	5	.8785	.8956	.8973	.8823	.8835
	7	.8764	.8768	.8941	.8693	.8788
	9	.8896	.8810	.8943	.8721	.8819
Sex DP	2	.1788	.1657	.1869	.1945	.1757
	3	.1805	.1657	.1869	.1969	.1757
	4	.1795	.1719	.1869	.1928	.1754
	5	.1660	.1663	.1919	.1921	.1461
	7	.1682	.1613	.1724	.2002	.1452
	9	.1795	.1583	.1659	.2040	.1469
Race DP	2	.2847	.2138	.1162	.1888	.2207
	3	.2962	.2138	.1162	.1888	.2207
	4	.2962	.2249	.1162	.1501	.2149
	5	.3343	.2305	.1269	.1611	.1917
	7	.3542	.1970	.1138	.1859	.1741
	9	.2481	.1970	.1126	.1933	.1974

SAFE-Stream (per fold, folds 1–5). Per-fold sex EO and race EO for SAFE-Stream are computed by the notebook (means and standard deviations in Table 2) but were not printed per fold .

A.2 Audited latent-group composition (Protocol B)

The audit draws 602 random samples per K (unseeded, so the samples differ across K). "Anchor" is the notebook's maximum single sex-or-race proportion. Female = sex code 0.

Table 5: Audited latent-group composition (Protocol B).

K	Group	Audited	Female	White	Black	Anchor
2	0	299	0.411	0.876	0.080	0.876
2	1	303	0.218	0.868	0.089	0.868
3	0	267	0.348	0.850	0.101	0.850
3	1	200	0.430	0.875	0.090	0.875
3	2	135	0.074	0.867	0.096	0.926
4	0	187	0.385	0.888	0.070	0.888
4	1	206	0.398	0.864	0.068	0.864
4	2	26	0.192	0.962	0.038	0.962
4	3	183	0.126	0.929	0.055	0.929
5	0	145	0.241	0.931	0.041	0.931
5	1	107	0.290	0.832	0.140	0.832
5	2	113	0.451	0.858	0.106	0.858
5	3	172	0.355	0.913	0.058	0.913
5	4	65	0.123	0.877	0.046	0.877
7	0	110	0.500	0.845	0.082	0.845
7	1	75	0.320	0.733	0.253	0.733
7	2	89	0.348	0.809	0.124	0.809
7	3	96	0.479	0.812	0.146	0.813
7	4	13	0.154	1.000	0.000	1.000
7	5	95	0.253	0.916	0.042	0.916
7	6	124	0.153	0.911	0.065	0.911
9	0	86	0.535	0.907	0.058	0.907
9	1	40	0.200	0.775	0.075	0.800
9	2	49	0.082	0.939	0.020	0.939
9	3	52	0.577	0.827	0.154	0.827
9	4	89	0.348	0.876	0.079	0.876
9	5	9	0.222	0.889	0.111	0.889
9	6	131	0.412	0.847	0.107	0.847
9	7	66	0.379	0.864	0.106	0.864
9	8	80	0.212	0.887	0.062	0.887