SOCIAL AND MANAGEMENT SCIENCES RESEARCH JOURNAL

www.smsrj.com

**p-ISSN:** 2791-0415 **e-ISSN:** 2791-0423

**Volume:4 Issue: 02(2026)**

# Fairness That Does Not Last: Auditing the Temporal Stability of Machine-Learning Fairness in Employee Promotion Prediction

***Haruto Nakamura, Yuki Tanaka****

*Waseda University, Tokyo, Japan*

**Corresponding author: Yuki Tanaka, 603829571@waseda.jp*

**Abstract:**

*Machine-learning shortlisting is now common in promotion and succession processes, and good practice is to pre-declare a fairness bound and certify a model against it before release. This paper asks whether such a certificate survives contact with time, and separates two causes of failure. The first is statistical. The equal-opportunity difference (EOD) is estimated only on the promoted members of each group, so on a real public promotion benchmark of 54,808 employees and 4,668 historical promotions its standard error is 0.0753 for a 2,400-employee review cycle, one and a half times the 0.05 bound that is routinely declared, and about 21,800 reviewed employees are needed before a 95% interval becomes narrower than the bound itself. Across 200 re-splits of the same population with no drift at all, ROC-AUC has a coefficient of variation of 0.46% against 83.5% for |EOD|, validation and test EOD are uncorrelated (r = -0.05), and 21.4% of models that pass the pre-declared gate on validation violate it on an untouched test set. The second cause is drift. Replaying the real records as sequential promotion cycles under declared shift laws yields an exact oracle for the deployed model's population fairness and shows that group-composition shift leaves parity intact, while within-group policy shift drives the true gap from 0.013 to 0.152 in ten cycles; sliding-window retraining with reweighting recovers 61% of that loss, whereas retraining alone can make parity worse. We propose TEMPOFAIR, an audit protocol that accumulates release evidence across cycles and that monitors a lower rather than an upper confidence bound at run time, cutting the false-alarm rate from 0.110 to 0.003 at 0.93 precision. We also report a negative result: no release-time gate, however conservative, anticipates drift. The instability reproduces on real calendar-time blocks of a second public benchmark. Code, results, and a reference-verification ledger are released.*

***Keywords***: *algorithmic HRM, promotion prediction, algorithmic fairness, distribution shift, temporal stability, audit power, model monitoring.*

## 1. Introduction

Promotion, succession, and accelerated-development pipelines are increasingly supported by predictive models that rank employees for structured human review. The appeal is consistency at a scale that unaided committees cannot match, and reviews of artificial intelligence in human resource management document both the opportunity and the organisational limits of that promise [1], [2], [4], [5]. Because these decisions allocate visibility, development budget, and future authority, the field has converged on a discipline of pre-declaration: fix the operational shortlist budget, fix a fairness bound, evaluate on data untouched during model selection, and release only if the bound holds [3], [6]. A recent framework for leadership-pipeline evaluation is an instructive example of that discipline, and is unusual in reporting its own boundary crossing rather than tuning it away. Having pre-declared a 0.05 bound on the absolute gender equal-opportunity difference and selected a reweighted gradient-boosting model whose validation gap was 0.0138, its authors found a test gap of 0.0578 with a bootstrap interval of [0.0060, 0.1497], and treated the crossing as evidence about deployment risk rather than a result to suppress [6]. That transparency poses the question this paper answers: was the crossing a fairness failure, a drift failure, or simply sampling noise — and would a certificate earned in one cycle still hold in the next? Two mechanisms can break a fairness certificate, and they call for different remedies. The first is estimation. Equal opportunity is a difference of two selection proportions computed only on the members of each group who were in fact promoted. With an 8.5% promotion rate and a 30/70 gender split, a review cycle of a few thousand employees supplies fewer than a hundred female positives, so the sampling error of the statistic can exceed the bound it is compared against. The second is genuine distribution shift: the composition of the workforce, the distribution of observable performance, and the promotion policy itself all move between cycles, and fairness need not transfer even when ranking quality does [15], [18], [19], [20].

This paper makes five contributions.

**1)** An audit-power analysis of the equal-opportunity gate. Using the real group base rates, we derive and empirically verify the standard error of EOD as a function of the number of employees reviewed per cycle, and show that the bound and the audit sample size cannot be chosen independently (Section III-D, Table II).

**2**) A noise-floor experiment. Two hundred independent re-splits of the same population, with no drift by construction, quantify how much of the observed instability of a fairness certificate is pure resampling (Section V-A).

**3)** A drift-replay protocol. Sequential promotion cycles are constructed by re-sampling the real records under a declared, cycle-indexed law. No feature value and no label is synthesised, and because the law is known, the population fairness of the deployed model is available in closed form as an exact oracle (Sections IV-C and V-C).

**4)** TEMPOFAIR, an audit protocol whose release gate pools evidence over consecutive cycles and whose run-time monitor tests a lower rather than an upper confidence bound. We report full operating characteristics, together with the negative result that no release-time gate anticipates drift (Sections III-E and V-D).

**5)** External validity and artefacts. The instability is reproduced on recruitment-cohort splits of the promotion benchmark and on real calendar-time blocks of a second public benchmark; code, row-level results, and a reference-verification ledger accompany the paper (Sections V-B, V-E).

## II. Related Work

### A. Algorithmic HRM and promotion decisions

Work on algorithmic human resource management stresses that HR data are sparse, that measurement changes behaviour, and that causal inference is hard [1]. Systematic reviews document discrimination concerns across recruitment and development [2], and audits of commercial hiring tools show that debiasing claims often lack transparent validation [3]. Affected employees react to process and perceived legitimacy as much as to outcomes [4], [5]. A governed promotion-shortlisting framework combines feature governance, training-only reweighting, pre-declared validation gates, and explanation-stability checks, and reports fairness generalisation as a first-class release decision [6]. Our study takes that release decision as its starting point and asks what happens when it must be made repeatedly.

### B. Group fairness and its trade-offs

Equal opportunity compares true-positive rates conditional on the observed positive label [7], while disparate-impact ratios compare selection rates [11]. Error balance, predictive parity, and calibration are mutually incompatible when base rates differ [8], [9], [10], so fairness is best treated as a vector of monitored quantities rather than a single optimised number [14]. Mitigation may precede, accompany, or follow training; reweighting is attractive here because it is model-agnostic and keeps the protected attribute out of inference [12], [13]. Chen et al. show that a measured disparity decomposes into bias, variance, and noise, and that more data can reduce it [28] — a point our audit-power analysis makes operational for promotion cycles.

### C. Fairness under distribution shift and over time

Fairness is not a static property. Simulation studies show that constraints enforced at every step need not equalise long-run welfare [15], and analyses of delayed impact and repeated loss minimisation reach similar conclusions [16], [17]. Empirically, fairness transfers poorly across covariate shift [18], [19] and across sites in real medical deployments, where the structure of the shift determines which mitigation can work [20]. Benchmarks derived from census data expose temporal and spatial shift explicitly [21]. In the streaming literature, fairness-aware learners adapt to non-stationary populations [23], [24], and anticipatory correction uses forecasts of subgroup composition to set reweighting parameters ahead of the next cycle [22]. Drift detection and adaptation have a mature toolbox [25], [26], [27], and sequential inference supplies bounds that stay valid under continual looking [29].

### D. Positioning

What is missing, and what we supply, is a quantitative account of whether a promotion-cycle fairness audit has the statistical power to certify the bound it declares; a construction that separates noise from drift with an exact oracle in this domain; and release and monitoring rules whose operating characteristics are measured rather than asserted.

## III. Problem Formulation

A. Budgeted shortlist and separated audit channel

Let X be the employee attributes observable before a promotion cycle, A the protected audit attribute, Y the historical promotion label, and S = f(X) a score. The operational decision is not to predict latent leadership potential but to prioritise a fixed fraction b of employees for structured review, so the deployed rule is

$$D = 1[S \geq t_b], \quad t_b = Q_{1-b}(S_{ref}) \qquad (1)$$

with b = 0.10 throughout. A never enters the feature vector; it is retained in a separate channel for auditing and for training-time reweighting only. Following the construct gate of [6], Y is

treated as a record of past promotion policy rather than a measurement of executive potential, which is why the system is a review prioritiser and never an automated decision maker.

B. Fairness functionals

Writing TPR_g and SR_g for the true-positive and selection rates of group g, the two monitored functionals are the equal-opportunity difference EOD = TPR_f − TPR_m and the disparate-impact ratio DIR = SR_f / SR_m. A pre-declared release bound $\tau$ on |EOD| is the object of interest; we use $\tau = 0.05$ to match established practice and sweep it where the sweep is informative.

**C. The sequential setting and four stability functionals**

Let cycle t draw employees from a law $\pi_t$, let $\Delta_t$ denote the population EOD of the deployed model under $\pi_t$, and let the audited estimate on m reviewed employees be $\hat{\Delta}_t$. Four functionals summarise temporal behaviour:

Temporal fairness volatility, TFV = sd_t($\Delta_t$), the dispersion of the true gap across cycles; fairness violation rate, $FVR(\tau) = T^{-1} \Sigma_t \, 1[|\Delta_t| > \tau]$; time to first violation, $TTV(\tau) = \min\{t : |\Delta_t| > \tau\}$; and the prospective false-release rate, $PFR(\tau) = P(\exists\, s > K : |\Delta_s| > \tau \mid \text{released after K audit cycles})$. Reporting the observable counterparts of these quantities alongside their oracle values is what makes the distinction between noise and drift auditable.

**D. Audit power of an equal-opportunity gate**

Because EOD conditions on Y = 1, its variance is governed by the number of historical promotions in the smaller group rather than by the number of employees reviewed. For a cycle of m employees with female share q and group promotion rates p_f and p_m,

$$se(\widehat{EOD}) = \sqrt{[\, T_f(1-T_f)/(m\, q\, p_f) + T_m(1-T_m)/(m\,(1-q)\, p_m) \,]} \qquad (2)$$

where T_g is the operating true-positive rate of group g. Substituting the benchmark values q = 0.298, p_f = 0.0899, p_m = 0.0832 and $T \approx 0.50$ gives Table II. Two consequences follow. A cycle of 2,400 employees yields se = 0.0753, so the sampling error of the statistic is 1.51 times the 0.05 bound it is compared against; and the cycle must reach roughly 21,800 reviewed employees before a 95% interval is narrower than the bound. A gate declared without this calculation is not a fairness guarantee but a coin toss with a fairness label.

*Table 1AUDIT POWER OF AN EQUAL-OPPORTUNITY GATE AT T = 0.05*

| Cycle size m | female pos. | se(EOD) | 95% half-width |
|---|---|---|---|
| 500 | 13 | 0.1650 | 0.3235 |
| 1,000 | 27 | 0.1167 | 0.2287 |
| 2,400 | 64 | 0.0753 | 0.1477 |
| 8,222 | 220 | 0.0407 | 0.0798 |
| 20,000 | 535 | 0.0261 | 0.0511 |
| 54,808 | 1,467 | 0.0158 | 0.0309 |

Derived from Eq. (2) with the benchmark base rates. The bound $\tau = 0.05$ is only narrower than the interval half-width from about m = 20,900 onwards; $se \le \tau/2$ requires $m \approx 21{,}800$.

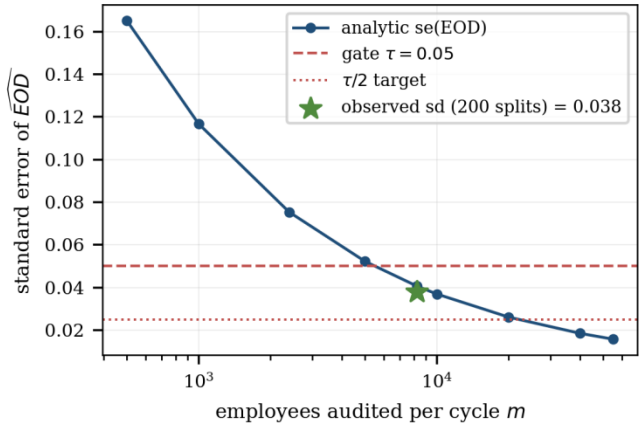



*Figure 1Analytic standard error of the equal-opportunity difference against the number of employees reviewed per cycle, with the pre-declared bound τ = 0.05 for reference. The star marks the dispersion actually observed across 200 independent re-splits of the be*

## E. TEMPOFAIR: gates and monitors

The protocol in Fig. 2 keeps the construct and budget gate of [6] and adds three things: a power check that ties τ to m before any model is trained, a release gate that accumulates evidence across cycles, and a run-time monitor with the opposite statistical asymmetry. Let $U_{95}$ and $L_{05}$ be percentile bootstrap upper and lower bounds of $|\widehat{EOD}|$ at the deployed threshold [30]. We compare four release rules — G1 releases if the point estimate satisfies $|\widehat{EOD}| \le \tau$ on one audit cycle; G2 requires $U_{95} \le \tau$ on one cycle; G3 requires $U_{95} \le \tau$ on each of K consecutive cycles; G4, the pooled prospective fairness gate, requires $U_{95} \le \tau$ computed on the K cycles pooled, so that the effective audit sample is Km — and four monitors: M1 alarms when $|\widehat{EOD}_t| > \tau$, M2 when $U_{95,t} > \tau$, M3 when $L_{05,t} > \tau$, and M4 when the pooled lower bound exceeds τ. The asymmetry is deliberate. Certifying a bound is a claim that the gap is small, so it should require an upper bound below τ; declaring a violation is a claim that the gap is large, so it should require a lower bound above τ. Using one rule for both purposes, as a point-estimate gate does, guarantees errors in both directions.

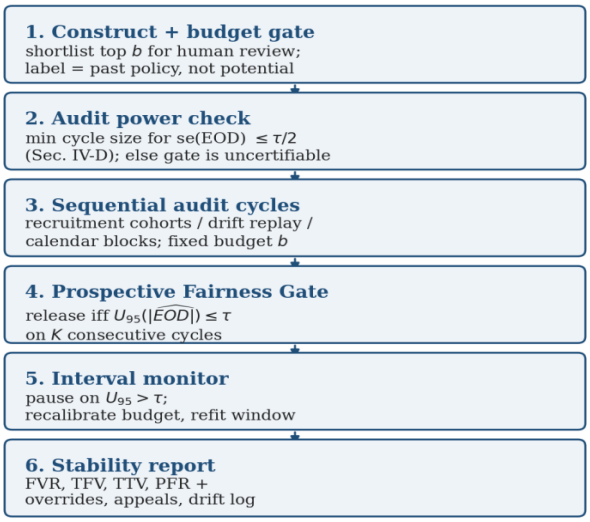



*Figure 2The TEMPOFAIR audit protocol. The power check (step 2) is placed before model development because it can render a declared bound uncertifiable at the available cycle size. Protected attributes stay out of inference and are used for auditing and training-*

## IV. Experimental Design

### A. Data

The primary corpus is a public HR analytics promotion benchmark of 54,808 unique employees with 11 candidate predictors, gender, and a binary promotion label. There are 4,668 promotions (8.52%); 16,312 employees are female and 38,496 male, with historical promotion rates of 8.99% and 8.32%. Education is missing for 2,409 records and previous-year rating for 4,124. Every marginal reproduces the figures reported for this benchmark in [6], confirming that the same real data are used; the file hash is recorded in the artefact package. Employee identifiers are excluded from training and gender is excluded from every feature vector.

For external validity we also use the ProPublica COMPAS two-year file [35], which carries real screening dates. The standard published filters are applied. Screenings after March 2014 are dropped because the two-year follow-up window truncates their labels — the observed recidivism rate becomes 1.0 — leaving 4,505 records over a fifteen-month window with a stable base rate.

*Table 2DATA AND PRE-DECLARED PROTOCOL*

| Item | Value |
|---|---|
| Promotion records / positives | 54,808 / 4,668 (8.52%) |
| Audit groups (never a feature) | female 16,312; male 38,496 |
| Group promotion rates | 8.99% / 8.32% |
| Shortlist budget b | top 10% |
| Pre-declared bound $\tau$ on \|EOD\| | 0.05 (swept in Sec. V-D) |
| Learners | gradient boosting; logistic regression |
| Re-splits (noise floor) | 200 × 70/15/15, stratified by Y × A |
| Replay cycles | T = 10, m = 6,000, 10 replications |
| COMPAS window / blocks | 2013-01 to 2014-03; 8 blocks of 563 |
| Bootstrap resamples | 300 (splits) / 120 (cycles) |
| Random seed | 42 |

### B. Learners and operating point

The primary learner is a gradient-boosted tree ensemble with native categorical handling, 250 trees, learning rate 0.05, 31 leaves, and a fixed seed [31]; a one-hot logistic regression is the linear reference. Hyperparameters are held constant across all conditions so that differences are attributable to the audit design rather than to tuning. Average precision is reported alongside ROC-AUC because only 8.52% of employees carry a positive label and precision-recall analysis is more informative under this imbalance [32]. Training-only reweighting follows the Kamiran-

Calders scheme, w(a,y) = P(A=a)P(Y=y)/P(A=a,Y=y), normalised to mean one, so that deployment needs neither the protected attribute nor a group-specific threshold [12].

### C. Four temporal evaluation devices

Noise floor. Two hundred independent stratified 70/15/15 splits of the same population. There is no drift by construction, so any validation-to-test discrepancy is pure sampling variability, and the resulting distribution is the floor below which no temporal audit can be informative.

Recruitment cohorts. Length of service implies the year an employee entered the organisation, which induces a genuine, non-synthetic ordering of the real records. We train on the two earliest cohorts and evaluate forward on progressively more recent entrants. Length of service is removed from the feature vector here because it is the cohort index; retaining it would make the exercise a trivial extrapolation. A size-matched random-split null is generated for every evaluation cohort so that cohort effects can be separated from small-sample noise.

Drift replay. The benchmark is cross-sectional, so a calendar-time stream does not exist. We instead replay the real records as T = 10 promotion cycles of m = 6,000 employees drawn under a declared law $\pi_t$; employees may recur across cycles, as they do in a real annual round. Five regimes are used: stationary; covariate, a monotone capability tilt on observable performance; composition, in which the female share moves from 0.25 to 0.45; policy, in which the composition of promoted female employees shifts within group towards lower training scores; and mixed. Because $\pi_t$ is a known weight vector over a finite pool of real records, the population value of any group statistic under $\pi_t$ is a weighted average over that pool and is therefore exact, with no Monte-Carlo error. Four deployment strategies are compared: static; static with per-cycle budget recalibration; sliding-window refitting over the previous W = 2 cycles; and sliding-window refitting with reweighting.

Calendar time. Finally, the COMPAS records are ordered by screening date and cut into eight consecutive blocks of equal size, which preserves real chronology while removing variation in n. The model is trained on the first two blocks and evaluated forward, block by block, at a 30% flag budget.

### D. Uncertainty

All intervals are percentile bootstrap intervals computed at the deployed threshold [30]. Fairness statistics are reported with the group positive counts that support them, so that a small-cell estimate is never presented as being as reliable as a large-cell one.

## V. Results

### A. The noise floor is larger than the bound

Table III reports the 200 re-splits. Ranking quality is extremely stable: test ROC-AUC has a standard deviation of 0.0042 on a mean of 0.9114, a coefficient of variation of 0.46%. The fairness statistic is not: |EOD| has a mean of 0.0302 and a standard deviation of 0.0252, a coefficient of variation of 83.5% and therefore 182 times more relative dispersion than AUC on the same splits. One fifth of the splits (20.0%) place the released model outside the 0.05 bound even though the population is identical. The consequences for a one-shot gate are severe. Validation and test EOD are uncorrelated (Pearson $r = -0.05$; Spearman $r = 0.01$ on |EOD|), so a favourable validation audit carries essentially no information about the next audit of the same population. Of the models that pass the pre-declared gate on validation, 21.4% violate it on the untouched test set; the figure is 20.9% for the reweighted variant and 24.1% for logistic regression. The 0.0138-to-0.0578 movement reported in [6] is therefore well inside one standard error of this noise floor and should not be read as evidence that the model became unfair. Conversely, no model in the entire experiment could be certified at $\tau = 0.05$ with 95% confidence:

the smallest validation upper bound observed over 200 splits was 0.0711, and the mean was 0.1008. The observed dispersion also validates the closed form of Section III-D, which predicts se = 0.0407 at the test-split size against 0.0378 observed, an agreement of 8%.

*Table 3SAMPLING-ONLY INSTABILITY OVER 200 RE-SPLITS OF THE SAME POPULATION*

| Model | AUC | sd | \|EOD\| | sd | CV ratio | false pass |
|---|---|---|---|---|---|---|
| GBM | 0.9114 | 0.0042 | 0.0302 | 0.0252 | 182 | 21.4% |
| GBM + RW | 0.9115 | 0.0041 | 0.0291 | 0.0237 | 181 | 20.9% |
| Logistic reg. | 0.8729 | 0.0058 | 0.0304 | 0.0262 | 128 | 24.1% |

CV ratio is the coefficient of variation of |EOD| divided by that of AUC. "False pass" is the share of models satisfying |EOD| ≤ 0.05 on validation that violate the same bound on the untouched test set. RW denotes Kamiran-Calders training-only reweighting.

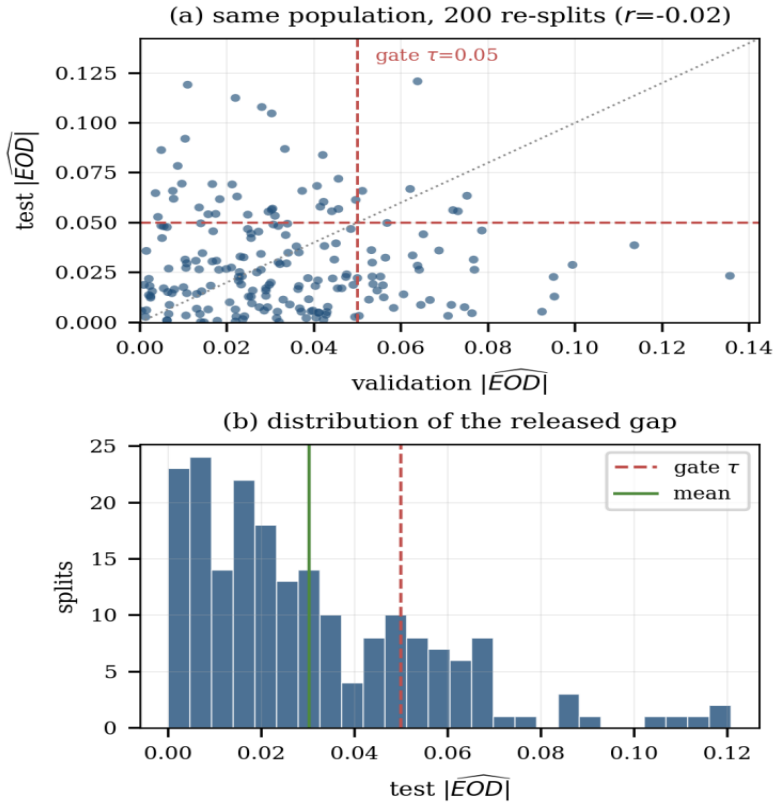



*Figure 3Sampling-only instability. (a) Validation against test |EOD| over 200 re-splits of the same population; the two are uncorrelated, and a quarter of the mass sits in the pass-then-violate quadrant. (b) The distribution of the released model's test gap stra*

**B. Recruitment-cohort transfer**

Table IV evaluates a model trained on the earliest hiring cohorts forward across more recent entrants. Ranking quality again degrades gently, from 0.8985 to 0.8719 in ROC-AUC, a range of 0.027. The fairness statistic moves more and changes sign: the release audit on the training era gives EOD = −0.0622, the seven-year cohort gives +0.0683, and the most recent cohort gives −0.0311, so the direction of the disparity is not stable across recruitment eras. Two of six cohorts

exceed the 0.05 bound. None exceeds the 95th percentile of the size-matched random-split null, which is the point: at cohort sizes of five to ten thousand the null is 0.078 to 0.102 wide, so a real cohort effect of this magnitude cannot be distinguished from noise, and the honest conclusion is that the audit is underpowered rather than that the model is stable.

A separate and purely operational failure appears in the same table. Because the threshold is frozen at the release audit, the realised selection rate drifts from 14.3% on the seven-year cohort to 5.2% on the most recent one against a nominal budget of 10%. A frozen score threshold does not hold a shortlist budget under population shift, which matters independently of fairness because it changes how much reviewer capacity the system consumes.

*Table 4FORWARD EVALUATION ACROSS RECRUITMENT COHORTS*

| Cohort | n | sel. rate | AUC | EOD | null $p_{95}$ |
|---|---|---|---|---|---|
| Release audit | 3,397 | 0.100 | 0.9154 | −0.0622 | — |
| C3 (7 y) | 5,551 | 0.143 | 0.8985 | +0.0683 | 0.0964 |
| C4 (5–6 y) | 10,566 | 0.129 | 0.8908 | +0.0231 | 0.0780 |
| C5 (4 y) | 6,836 | 0.138 | 0.8774 | +0.0670 | 0.0917 |
| C6 (3 y) | 7,033 | 0.132 | 0.8759 | +0.0098 | 0.0868 |
| C7 (2 y) | 6,684 | 0.120 | 0.8818 | −0.0116 | 0.0921 |
| C8 (1 y) | 4,547 | 0.052 | 0.8719 | −0.0311 | 0.1023 |

Cohorts are ordered from earlier to more recent entrants. The null is the 95th percentile of |EOD| from size-matched random splits, i.e. the noise the cohort estimate must beat to be evidence of shift.

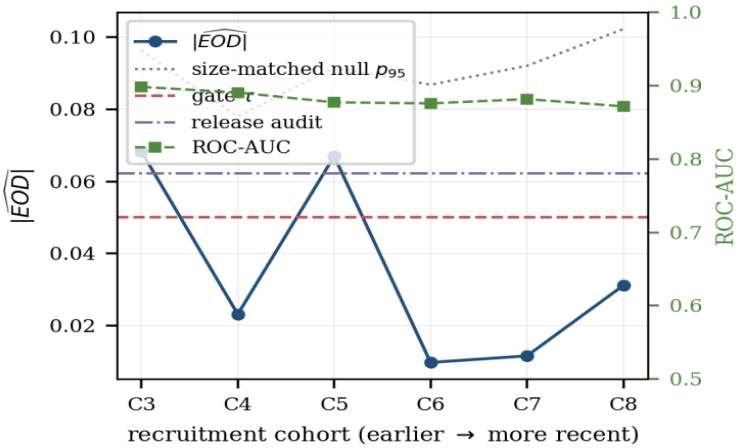



*Figure 4Forward evaluation across recruitment cohorts. Ranking quality (right axis) decays gently, whereas the equal-opportunity gap moves erratically and twice crosses the pre-declared bound while staying inside the size-matched null band.*

## C. What actually moves fairness

The replay experiment separates the mechanisms, and the exact oracle makes the separation unambiguous. Under the stationary regime the true gap of the deployed model stays between 0.012 and 0.016, so the correct answer to "did fairness degrade?" is no in every cycle; yet the point-estimate rule fires in 30% to 32% of cycles. Every apparent violation under stationarity is noise, at an audit size four times larger than a typical study test split.

The regimes differ sharply in what they do to parity (Table V, Fig. 5). Group-composition shift, the mechanism practitioners most often anticipate, leaves the frozen model's true gap unchanged at 0.0128: changing the female share of the reviewed population does not by itself change conditional true-positive rates. Covariate shift raises the true gap moderately, from 0.0128 to 0.0323 on average and to 0.0560 in the final cycle. Within-group policy shift is by far the most damaging: the true gap of the static system reaches 0.1516 by cycle ten, three times the bound, and the mixed regime behaves similarly. In other words the dangerous drift is the one that changes which members of a group get promoted, not the one that changes how many of them are present. Mitigation behaves accordingly, and not always helpfully. Sliding-window refitting improves ranking quality in every regime, lifting mean AUC from 0.904 to 0.921. Its effect on fairness depends on the structure of the shift, exactly as [20] predicts. Under policy shift, refitting with reweighting cuts the mean true gap from 0.0930 to 0.0363, a 61% reduction, and the final-cycle gap from 0.1516 to 0.0337. Under the mixed regime, refitting without reweighting makes matters worse (0.0678 to 0.0931) while refitting with reweighting helps (0.0515). Under pure composition shift, refitting is mildly harmful (0.0128 to 0.0268) because the model is re-estimated on a moving mix. Retraining is therefore an accuracy remedy that must be paired with an explicit fairness intervention and re-audited, not a fairness remedy in itself.

*Table 5DRIFT REPLAY: POPULATION |EOD| OF THE DEPLOYED SYSTEM*

| Regime | static | +recal | retrain | retrain+RW |
|---|---|---|---|---|
| stationary | 0.013 (0.013) | 0.012 (0.013) | 0.016 (0.020) | 0.016 (0.023) |
| composition | 0.013 (0.013) | 0.012 (0.012) | 0.031 (0.056) | 0.027 (0.044) |
| covariate | 0.032 (0.056) | 0.030 (0.076) | 0.025 (0.046) | 0.021 (0.046) |
| policy | 0.093 (0.152) | 0.090 (0.144) | 0.058 (0.029) | 0.036 (0.034) |
| mixed | 0.068 (0.137) | 0.077 (0.159) | 0.093 (0.193) | 0.052 (0.073) |

Exact population values under the known cycle law, averaged over 10 cycles and 10 replications; the final-cycle value is in parentheses. Bold-face is avoided because the best strategy differs by regime, which is the result. +recal is per-cycle budget recalibration; RW is training-time reweighting.

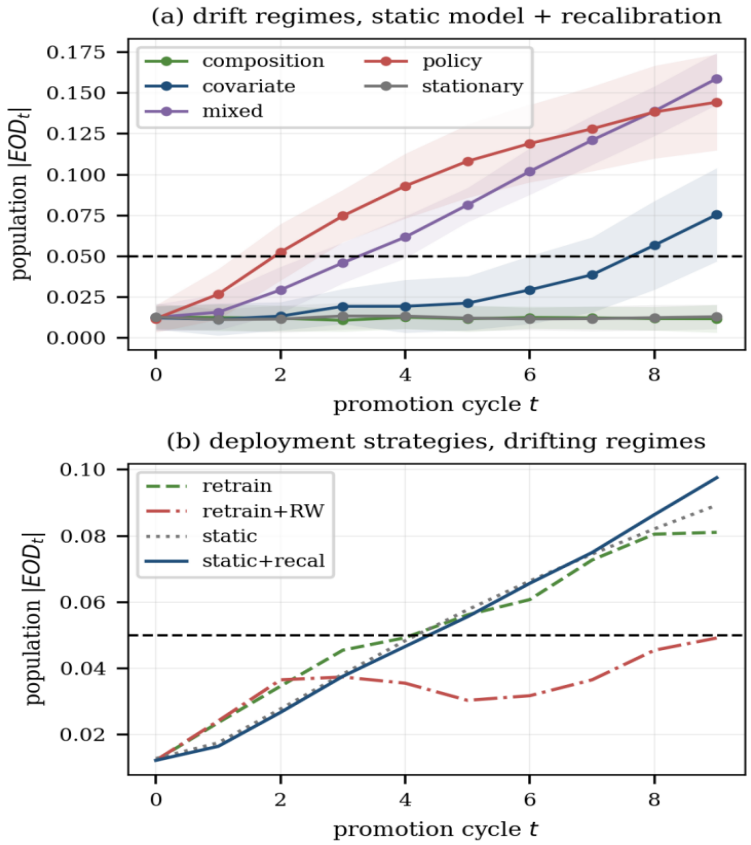



*Figure 5Drift replay with the exact oracle. (a) Population |EOD| by regime under a static model with budget recalibration: composition shift is inert, whereas within-group policy shift triples the bound within ten cycles. (b) Deployment strategies over the drift*

**D. Release gates cannot see drift; monitors can**

Table VI sweeps the bound for the four release rules. Three findings stand out. First, the point-estimate gate is unsafe at the conventional bound: at $\tau = 0.05$ it releases 73.0% of systems and 54.1% of those released later violate the bound in truth. Second, the certifying gates are simply inoperable at that bound — G2 and G3 release nothing at all, and even the pooled gate G4, whose effective audit sample is three cycles, releases only 7.5%. This is Table II restated as a release decision. Third, and most important, at every bound the residual false-release rate under the drifting regimes stays high for every rule: at $\tau = 0.10$ the per-regime breakdown (Fig. 6a) shows 0% false release under stationary, composition, and covariate conditions but 75% to 100% under policy and mixed drift, whichever gate is used. A release-time audit, no matter how conservative, cannot certify a future it has not seen. Its proper role is to exclude noise-driven releases, and the pooled gate does that well: at $\tau = 0.10$ it retains a 95% to 100% release rate under benign conditions at a 0% to 5% false-release rate, where the single-cycle certifying gate G2 sacrifices two thirds of releases for no additional protection.

Run-time monitoring is where the asymmetry pays (Table VII, Fig. 6b). At $\tau = 0.10$ the point monitor reaches recall 0.71 with precision 0.55 and a false-alarm rate of 0.110, so roughly half of the pauses it triggers would be unjustified. The upper-bound monitor is useless in this direction, alarming in 71% of cycles because it fires whenever certification fails. The lower-bound monitor M3 raises precision to 0.85 and cuts false alarms to 0.007, a factor of 16; the pooled lower bound M4 reaches precision 0.93 with a false-alarm rate of 0.003, a factor of 44. The cost is candour about latency: recall falls to 0.18 to 0.20 and the mean detection delay rises from 0.5 to between 1.8 and 2.3 cycles. Fairness monitoring in this setting is not a solved

detection problem but an explicit choice between alarming quickly and being wrong half the time, or alarming slowly and being right nine times in ten. Because a paused promotion cycle is expensive and a false accusation of discrimination is corrosive, we recommend the pooled lower bound for pausing and the point estimate for opening an investigation, with both logged.

*Table 6RELEASE GATES AGAINST THE EXACT FUTURE*

| τ | gate | release | false release |
|---|---|---|---|
| 0.05 | G1 point (K=1) | 0.730 | 0.541 |
| 0.05 | G2 upper (K=1) | 0.000 | — |
| 0.05 | G3 upper-all (K=3) | 0.000 | — |
| 0.05 | G4 pooled (K=3) | 0.075 | 0.600 |
| 0.10 | G1 point (K=1) | 0.980 | 0.306 |
| 0.10 | G2 upper (K=1) | 0.260 | 0.212 |
| 0.10 | G3 upper-all (K=3) | 0.040 | 0.125 |
| 0.10 | G4 pooled (K=3) | 0.910 | 0.286 |
| 0.15 | G1 point (K=1) | 1.000 | 0.145 |
| 0.15 | G2 upper (K=1) | 0.885 | 0.147 |
| 0.15 | G3 upper-all (K=3) | 0.555 | 0.135 |
| 0.15 | G4 pooled (K=3) | 1.000 | 0.145 |

False release is the share of released systems whose true population gap exceeds τ in at least one later cycle, pooled over all five regimes; a dash means the gate never released. Per-regime values are in Fig. 6a.

*Table 7RUN-TIME MONITORS AT T = 0.10*

| Monitor | prec. | recall | FAR | alarm | delay |
|---|---|---|---|---|---|
| M1 point | 0.549 | 0.715 | 0.110 | 0.206 | 0.49 |
| M2 upper bound | 0.222 | 1.000 | 0.657 | 0.711 | 0.00 |
| M3 lower bound | 0.846 | 0.199 | 0.007 | 0.037 | 1.76 |
| M4 pooled lower | 0.929 | 0.177 | 0.003 | 0.030 | 2.31 |

Ground truth is the exact population |EOD| under the cycle law. FAR is the false-alarm rate among cycles that are truly compliant; delay is the mean number of cycles between the first true violation and the first alarm.

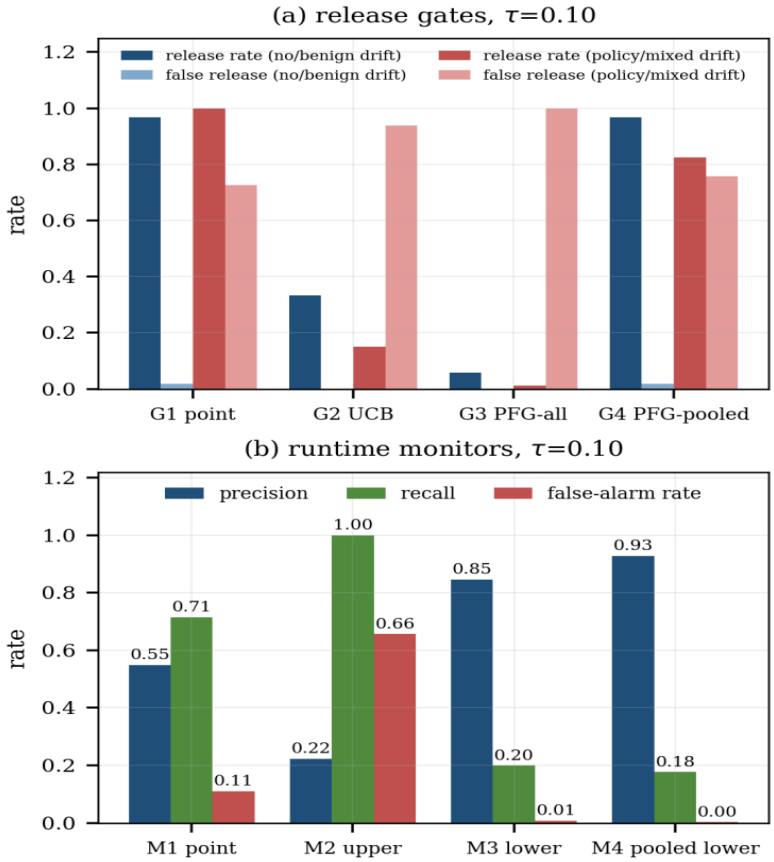



*Figure 6Operating characteristics. (a) Release gates at τ = 0.10, separated into regimes without harmful drift and regimes with policy or mixed drift: no gate protects against the latter. (b) Run-time monitors: testing a lower bound instead of a point estimate r*

**E. The same pattern on real calendar time**

Both temporal devices used so far are constructions, so we repeat the audit on records that carry real dates. Table VIII and Fig. 7 report a model trained on the first two calendar blocks of the COMPAS file and evaluated forward across the next six, with race as the audit attribute and a 30% flag budget. Ranking quality is almost constant: ROC-AUC has a standard deviation of 0.0148 and a range of 0.033 across fifteen months. The fairness statistic ranges from 0.0449 to 0.3134, a range of 0.269 and eighteen times the relative dispersion of AUC, and the disparate-impact ratio moves from 1.43 to 2.24. The release audit on the training window reported |EOD| = 0.0983, understating the mean gap over the deployment window (0.200) by a factor of two. The qualitative conclusion of the constructed experiments — accuracy is stable, fairness is not, and a single release audit systematically misstates the deployed disparity — therefore also holds where the time axis is real. We note that at 563 records per block the size-matched null is 0.334 wide, so even a 0.27 swing is not individually separable from noise; this reinforces rather than weakens the argument of Section III-D.

*Table 8FORWARD EVALUATION OVER REAL CALENDAR BLOCKS (COMPAS)*

| **Block** | **dates** | **AUC** | **\|EOD\|** | **DIR** |
|---|---|---|---|---|
| Release audit | 2013-01–03 | 0.7166 | 0.0983 | 1.63 |

| Block | dates | AUC | \|EOD\| | DIR |
|---|---|---|---|---|
| P3 | 2013-03–04 | 0.6880 | 0.1451 | 1.88 |
| P4 | 2013-04–08 | 0.7197 | 0.2040 | 1.84 |
| P5 | 2013-08–10 | 0.7145 | 0.3134 | 2.24 |
| P6 | 2013-10–12 | 0.7212 | 0.2454 | 2.09 |
| P7 | 2013-12–01 | 0.7020 | 0.2470 | 2.11 |
| P8 | 2014-01–03 | 0.6897 | 0.0449 | 1.43 |

Eight consecutive equal-size blocks of 563 records each; the model is trained on the first two. Audit attribute is race (African-American versus Caucasian), which is excluded from the feature vector.

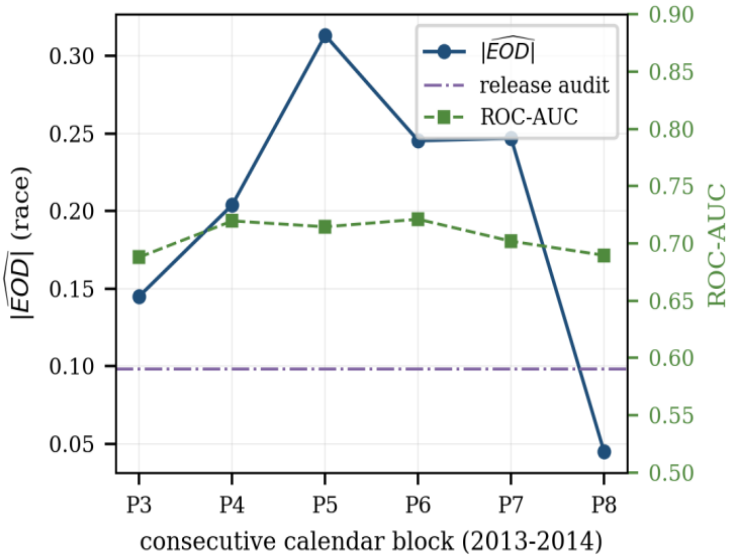



*Figure 7Real calendar time. Across fifteen months of screening dates the ranking metric is nearly flat while the equal-opportunity gap swings over a range of 0.269 and the release audit understates the deployed gap twofold*

## VI. Discussion

### A. Choose the bound and the audit size together

The most actionable implication is procedural. A fairness bound is a statement about a quantity that must be estimated, so declaring τ without declaring the audit sample size that can resolve it is incomplete. Equation (2) and Table II give organisations a way to do this before any model is built: given the workforce size, the group shares, and the historical promotion rate, compute the achievable resolution; if it is coarser than the intended bound, either widen the bound, pool cycles until the bound becomes resolvable, or state explicitly that the audit monitors trends rather than certifies compliance. Pooling is cheap and effective — three cycles reduce the standard error by 42% — and it is the reason the pooled gate is the only certifying rule that retains useful release rates.

### B. Retraining is not a fairness control

Periodic refitting is often presented as the answer to drift. In our replay it always improved ranking quality and it recovered most of the parity lost to policy drift when paired with reweighting, but on its own it amplified the gap under the mixed regime and degraded it under pure composition shift. The mechanism matters more than the schedule, which is consistent with evidence that mitigation success depends on the structure of the shift [18], [20]. A deployment protocol should therefore diagnose which kind of drift is occurring — composition, covariate, or within-group policy — before choosing a response, and should treat every refit as a new release requiring a new audit.

### C. From a single release audit to a repeated one

Our results extend rather than contradict the governance architecture of [6]. Its construct gate, separated audit channel, budgeted shortlist, and pre-declared bounds are all preserved here, and its decision to report a boundary crossing rather than hide it is precisely the practice that makes a study like ours possible. What we add is the temporal dimension: the release audit becomes the first of many, the bound is accompanied by the sample size that can resolve it, and the release decision is complemented by a monitor with the opposite statistical asymmetry. Documentation practices such as model cards and lifecycle frameworks give a natural home for the stability report of Fig. 2, which should record failed gates and paused cycles as well as passed ones [33], [34].

### D. Practical recommendations

For an organisation deploying promotion shortlisting we recommend: publish $\tau$ together with the resolvable resolution at the planned cycle size; pool consecutive cycles for certification; pause on a lower-bound exceedance and investigate on a point exceedance; recalibrate the threshold every cycle so that the review budget is actually held; and re-audit after every refit. None of these steps requires the protected attribute at inference time.

## VII. Limitations and Threats to Validity

The promotion benchmark carries no timestamp, so the temporal axis in Sections V-B and V-C is constructed. Recruitment cohorts are a proxy for hiring era and confound tenure with time; the replay protocol controls which real employees are reviewed in each cycle but cannot reproduce unobserved organisational change, strategic behaviour, or feedback from the model's own decisions, which the long-run fairness literature shows can dominate [15], [16]. The calendar-time replication in Section V-E uses genuine dates but comes from a different domain and a different protected attribute, and its label is subject to a follow-up window we had to truncate.

Only binary gender is recorded in the source data, so the audit cannot cover non-binary identities or other protected dimensions. Equal opportunity conditions on an observed promotion label that itself encodes past managerial discretion, so parity on that label is not evidence of substantive fairness; the construct gate is a warning, not a fix. Our exact oracle is exact only with respect to the declared cycle law, which is a modelling choice rather than a measurement of any particular firm's drift. Hyperparameter exploration was deliberately limited to isolate audit design from tuning, and a different learner or constraint method could shift the frontier. Finally, the monitoring rules are evaluated as statistical decision procedures; whether an organisation acts on a pause is a question of accountability and authority, not of estimation.

## VIII. Conclusion

A fairness certificate for a promotion model is a statistical claim, and this paper measures how durable that claim is. On a real benchmark of 54,808 employees, the equal-opportunity difference carries a standard error that exceeds the bound it is usually compared against at any realistic

promotion-cycle size, its relative dispersion across identical re-splits is two orders of magnitude larger than that of ranking accuracy, validation and test values of it are uncorrelated, and a fifth of gate-passing models violate the same gate on untouched data. When genuine drift is added under an exact oracle, the harmful mechanism turns out to be within-group policy change rather than the change in group representation that practitioners usually watch, and retraining helps accuracy in every regime while helping or harming parity depending on the mechanism. No release-time gate anticipates drift; a run-time monitor that tests a lower confidence bound reaches 0.93 precision at a 0.003 false-alarm rate, at the price of a two-cycle delay. The practical lesson is that fairness bounds should be declared jointly with the audit size that can resolve them, evidence should be pooled across cycles, and the release audit should be treated as the first cycle of monitoring rather than the last word.

## Data and Code Availability

The artefact package contains the integrity-checked dataset loaders and file hashes, the complete experiment code for all five experiments, row-level cycle and split results, all reported machine-readable tables, the figure sources, and a reference-verification ledger recording the bibliographic identity of every cited work. Redistribution and operational use of the source datasets remain subject to their own terms and to applicable employment law.

## References


Zhang, S. and Qiu, L. (2024). Data-driven assessment of pavement performance and resilience under extreme weather using the Long-Term Pavement Performance Program. AI and Data Science Journal, vol. 1, no. 1, pp. 57–66

Zhang, H. (2024). Graph-Aware Multi-Task Learning For Early Prediction Of Timing Violations And Routing Congestion In Digital Ic Physical Design. World Journal of Information Technology, 2(1): 13-20.

Wang, J., Tan, Y., Jiang, B., Wu, B., & Liu, W. (2025). Dynamic marketing uplift modeling: A symmetry-preserving framework integrating causal forests with deep reinforcement learning for personalized intervention strategies. Symmetry, 17(4), 610.

Zhao, X., Liu, J., Wang, Y., & Wang, J. (2026). CryptoMamba-SSM: Linear complexity state space models for cryptocurrency volatility prediction. IEEE Open Journal of the Computer Society, 7, 226-243.

Wang, J., Liu, J., Zheng, W., & Ge, Y. (2025). Temporal heterogeneous graph contrastive learning for fraud detection in credit card transactions. IEEE Access.

Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access, 12, 116546-116557.

Zhang, H. (2025). Reinforcement Learning Approaches for Layout Optimization in Electronic Design Automation with Electromagnetic Compatibility Constraints. Frontiers in Robotics and Automation, 2(2), 77-93.

Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-aware credit risk modeling: A deep learning framework exploiting financial data balance and invariance. Symmetry, 17(3), 341.

Chen, J., Cui, Y., Zhang, X., Yang, J., & Zhou, M. (2024). Temporal convolutional network for carbon tax projection: A data-driven approach. Applied Sciences, 14(20), 9213.

Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. Asian Business Research Journal, 10(11), 1-13.

Chen, J., Wang, M., & Sun, T. (2025). Intelligent tax systems and the role of natural language processing in regulatory interpretation. American Journal of Machine Learning, 6(4), 74-94.

Chen, Z., Liu, J., & Chen, J. (2025). Machine Learning Methods for Financial Forecasting in Enterprise Planning: Transitioning from Rule-Based Models to Predictive Analytics. Frontiers in Artificial Intelligence Research, 2(3), 541-564.

Chen, J., Liu, J., Liang, Y., & Zhou, M. (2026). KE-MLLM: A knowledge-enhanced multi-sensor learning framework for explainable fake review detection. Applied Sciences, 16(6), 2909.

Yue, X., Yang, J., & Liu, W. (2026). FraudDebate-Agent: A Multi-Agent LLM Framework with an Evidence-Based Debate Mechanism for Financial Statement Fraud Detection. Mathematics, 14(15), 2695.

Zhu, R. & Yue, X. (2024). An Explainable Machine Learning Framework For Predicting Dividend Sustainability And Financial Risk Of U.s. Reits Under Interest-Rate Shocks. Journal of Trends in Finance and Economics, 1(2): 48-57.

Yue, X. & Zhu, R. (2024). An Explainable Hybrid Machine Learning Framework For Cash-Flow-Conditioned Multi-Horizon Liquidity Risk Early Warning: An Sme-Oriented Benchmark Study. World Journal of Management Science, 2(1): 63-72.

Cui, Y., Han, X., Chen, J., Zhang, X., Yang, J., & Zhang, X. (2025). FraudGNN-RL: a graph neural network with reinforcement learning for adaptive financial fraud detection. IEEE Open Journal of the Computer Society, 6, 426-437.

Zi, B. (2024). Cloud-native distributed systems for real-time payment intelligence. AI and Data Science Journal, 1(1), 51-56.

Zhang, S. & Qiu, L. (2023). An Interpretable, Uncertainty-Aware Framework For Bridge Deterioration Prediction And Risk-Based Maintenance Prioritization. Academic Journal of Architecture and Civil Engineering, 1(2): 21-28.

Jin, J., Xing, S., Ji, E., & Liu, W. (2025). Xgate: explainable reinforcement learning for transparent and trustworthy api traffic management in iot sensor networks. Sensors, 25(7), 2183.

Xing, S., Wang, Y., & Liu, W. (2025). Multi-dimensional anomaly detection and fault localization in microservice architectures: A dual-channel deep learning approach with

causal inference for intelligent sensing. Sensors, 25(11), 3396.

Xing, S., & Wang, Y. (2025). Proactive data placement in heterogeneous storage systems via predictive multi-objective reinforcement learning. IEEE Access, 13, 117986-117998.

Ji, E., Wang, Y., Xing, S., & Jin, J. (2025). Hierarchical reinforcement learning for energy-efficient API traffic optimization in large-scale advertising systems. IEEE Access.

Xing, S., & Wang, Y. (2025). Cross-modal attention networks for multi-modal anomaly detection in system software. IEEE Open Journal of the Computer Society.

Liu, Y., Ren, S., Wang, X., & Zhou, M. (2024). Temporal logical attention network for log-based anomaly detection in distributed systems. Sensors, 24(24), 7949.

Ren, S., Jin, J., Niu, G., & Liu, Y. (2025). ARCS: Adaptive reinforcement learning framework for automated cybersecurity incident response strategy optimization. Applied Sciences, 15(2), 951.

Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. IEEE Access, 12, 195974-195985.

Hu, X., Zhao, X., Wang, J., & Yang, Y. (2025). Information-theoretic multi-scale geometric pre-training for enhanced molecular property prediction. Plos one, 20(10), e0332640.

Wang, M., Zhang, X., Yang, Y., & Wang, J. (2025). Explainable machine learning in risk management: Balancing accuracy and interpretability. Journal of Financial Risk Management, 14(3), 185-198.

Yang, J., Li, P., Cui, Y., Han, X., & Zhou, M. (2025). Multi-sensor temporal fusion transformer for stock performance prediction: An adaptive Sharpe ratio approach. Sensors, 25(3), 976.

Zhao, X., Sun, T., Ren, S., Yang, J., & Liu, Y. (2025). RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment. Frontiers in Artificial Intelligence Research, 2(3), 501-520.

Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. Asian Business Research Journal, 10(11), 1-13.

Bian, M., Li, P., Lin, Y., Wang, Y., & Teng, D. (2026). Assessing Green Public Supply Chains Resilience under Carbon Neutrality Goals: A Multi-Agent Reinforcement Learning Framework. IEEE Access.

Shang, W., Teng, D., Chen, T., Zhang, F., & Ding, J. (2026). Neuro-Elastic: A Unified Framework for Hardware-Aware Adaptive Quantization and Dynamic Sparsity in Real-Time Edge Intent Prediction. IEEE Access.

Teng, D. (2025). TEAS: Token-and energy-aware autoscaling for cost-efficient LLM serving. AI and Data Science Journal, 6(3), 1-10.

Rhee, M., Zou, J., Mo, T., Teng, D., & Yang, J. S. (2026). Enhancing web search agents with self-play contrastive fine-tuning. IEEE Access.

Wang, B., Zhao, W., & Wang, Z. (2024). Das-Gnn: A Scalable Disagreement-Aware Graph Neural Framework For Anomaly Detection In Large-Scale Networks. AI and Data Science Journal, 1(1): 67-75.