DRIFT-CONDITIONED VALUE OF INFORMATION FOR COSTLY MULTI-LABEL STREAMS WITH DELAYED LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cost-aware prediction in a multi-label stream must decide which measurement to purchase before delayed labels reveal which risks have changed. Aggregate stream state is insufficient when drift is label local and candidate features have label-specific utility. We formulate *drift-conditioned value of information*: a delayed residual process estimates a label-local drift posterior, and each feature is scored by its expected reduction in class-balanced decision risk per unit cost, weighted by that posterior. The resulting policy reallocates budget when and where drift aligns with feature-specific risk reduction. On the chronological MultiEURLEX stream, DRIFTBUDGET achieves 70.4 macro-F1 and 58.0 tail recall at 44.5 normalized cost, compared with 66.1, 50.2, and 46.5 for DiFA, while reducing recovery from 17.9 to 10.8 steps. Posterior, global-state, delayed-credit, and delay controls at comparable cost support label-local drift conditioning as the mechanism behind the changed acquisition policy.

1 INTRODUCTION

A deployed predictor may observe a cheap feature subset first and purchase additional measurements only before a deadline. Delayed labels make this decision stale: the policy must act before the errors most relevant to the current item are observable. In multi-label tasks the mismatch is more structured. One label can change while aggregate loss moves little, and the feature that repairs that label may be irrelevant to the others. A global response can therefore spend the same budget on the wrong measurements.

Active feature acquisition estimates the value of optional measurements under missingness or explicit cost. EDDI uses latent expected information gain (Ma et al., 2019), DiFA learns differentiable cost-aware acquisition (Ghosh & Lan, 2023), explainability-driven AFA learns instance-specific rankings (Guney et al., 2025), and online active-feature trees adapt decisions in streams (Rahbar et al., 2023). Retrospective evaluation additionally requires explicit identification and support assumptions because deploying a new acquisition policy changes the observed-data distribution (von Kleist et al., 2025). Chronological evaluation exposes temporal generalization failures, while label-wise analysis shows that aggregate metrics can obscure heterogeneous label behavior. The missing decision variable is not merely whether the stream shifted, but whether a candidate feature can reduce the risks of the labels for which delayed evidence indicates change.

Central distinction. Drift changes an acquisition decision only through the interaction between a label-local drift state and the label-specific risk reduction of a candidate feature; a global drift alarm cannot represent this ranking.

This yields three contributions:

1. We formulate delayed multi-label acquisition as label-local, drift-conditioned value of information, making the acquisition target an expected reduction in class-balanced risk rather than generic uncertainty.

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.

2. We couple a delayed residual posterior to label-specific feature utility in one cost-normalized score, yielding explicit equivalence and failure predictions for global shift, misspecification, and inadequate support.
3. We test the proposed interaction with matched-cost policy comparisons and posterior, global-state, delayed-credit, and delay controls on the reported chronological stream.

2 Related Work and Novelty Boundary

Cost-aware acquisition and evaluation. Latent information gain, differentiable acquisition, explainability-driven ranking, missing-data modeling, and online feature acquisition provide established ways to value or reconstruct optional measurements (Ma et al., 2019; Cao et al., 2018; Ghosh & Lan, 2023; Guney et al., 2025). Their deployment-time value estimates are primarily instance- or state-conditioned. Retrospective policy evaluation is a separate problem because a changed acquisition policy changes what is observed; causal interpretation therefore requires explicit identification and support assumptions (von Kleist et al., 2025).

Delayed and label-wise adaptation. Delayed online learning and delayed-feedback models explain how supervision arrives after a decision (Joulani et al., 2013; Chapelle, 2014), while chronological and label-wise studies expose temporal and heterogeneous-label failures (T.y.s.s et al., 2024; Chalkidis & Søgaard, 2022). The novelty claim is not drift detection, cost normalization, or delayed updating in isolation. It is the decision-facing interaction between a delayed label-local state and the label-specific risk reduction of each candidate feature. A global detector followed by DiFA is therefore the nearest mechanism control: if it matches DriftBudget under heterogeneous drift, the proposed interaction has not added useful information.

3 Problem Formulation

At event time t, an item has optional features $x_t \in \mathbb{R}^m$, target $y_t \in \{0,1\}^L$, feature costs $c \in \mathbb{R}^m_+$, and label delay $d_t \geq 0$. The policy observes a mandatory subset S_t^0 and history $\mathcal{H}_t$ containing only labels whose arrival time is at most t. It sequentially chooses $S_t \supseteq S_t^0$ with $\sum_{j \in S_t \setminus S_t^0} c_j \leq B$. For class-balanced label loss ℓ_ℓ, the deployment objective is the risk–cost frontier induced by

$$\min_{\theta,\pi} \mathbb{E}\left[\sum_t \sum_{\ell=1}^{L} w_\ell \ell_\ell(f_\theta(x_{t,S_t}), y_{t\ell}) + \lambda \sum_{j \in S_t \setminus S_t^0} c_j\right]. \tag{1}$$

Neither the current label nor a future drift boundary is available when S_t is chosen. Costs and delays imposed on a retrospective dataset are declared interventions, not naturally observed metadata. Counterfactual acquisition value is interpreted causally only under the registered no-direct-effect, no-unobserved-confounding, and support conditions; otherwise it is a model-based diagnostic (von Kleist et al., 2025).

4 Drift-Conditioned Value of Information

4.1 Delayed label-local state

For label ℓ, a calibrated residual $e_{s\ell}$ is created only when $y_{s\ell}$ arrives. A state-space filter estimates

$$r_{t\ell} = \mathbb{P}(z_{t\ell} = 1 \mid \{e_{s\ell} : s + d_s \leq t\}), \tag{2}$$

where $z_{t\ell}$ denotes a changed conditional-risk regime. Between arrivals the posterior is propagated without using future labels. Tail-label shrinkage toward a shared prior is a supporting estimation choice, not a separate contribution, and a purely global state is retained as a required ablation.

4.2 Risk-based conditional feature value

For candidate feature $j \notin S$, define the expected reduction in label-specific decision risk

$$\Delta_{t\ell}(j \mid S) = \mathcal{R}_{t\ell}(S) - \mathbb{E}_{x_j \mid x_{t,S}, \mathcal{H}_t}[\mathcal{R}_{t\ell}(S \cup \{j\})], \tag{3}$$

where $\mathcal{R}_{t\ell}$ is calibrated class-balanced predictive risk. The core score is

$$V_t(j \mid S) = \frac{\sum_\ell w_\ell[\beta + (1-\beta)r_{t\ell}]\,\Delta_{t\ell}(j \mid S)}{c_j + \varepsilon}. \tag{4}$$

The policy greedily acquires the feasible positive-value feature and stops below a validation-fixed threshold. The residual filter, conditional feature model, and greedy solver implement Eq. 4; they are supporting components rather than parallel novelty claims. When a delayed label arrives, a timestamped decision log assigns the realized label-wise risk reduction back to the acquisition event and updates the feature-value model; the "w/o delayed credit" ablation omits this assignment while retaining the delayed drift-state update. No policy-gradient trace or separate reinforcement-learning mechanism is used.

4.3 Mechanism predictions

When drift evidence is stationary or equal across labels and the remaining controller inputs match, $r_{t\ell}$ provides no useful label ranking; the label-local term therefore predicts no advantage over a global-state control. When the conditional feature model is wrong, $\Delta_{t\ell}$ can have the wrong sign even with perfect state estimation. When labels are too delayed or too rare to localize change, the posterior should remain diffuse and the policy should revert toward its shared prior. These predictions connect the problem, mechanism, and observable diagnostics.

Figure 1 makes the two information paths explicit: only arrived residuals update the label-local state, whereas the conditional feature model estimates what an unobserved measurement would change. Their product in Eq. 4, rather than either branch alone, determines the acquisition ranking.

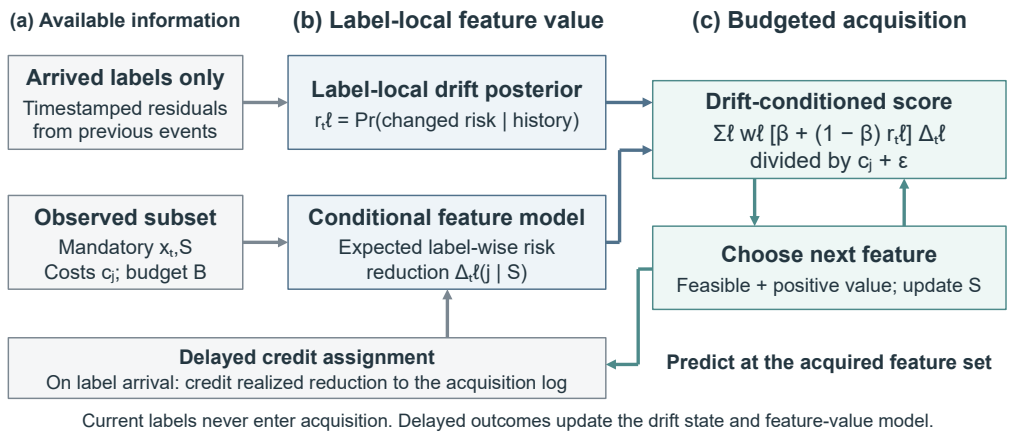

Figure 1: Method overview. Delayed label-local drift posteriors weight label-specific expected decision-risk reduction from the conditional feature model; the cost-normalized score drives a budgeted positive-value acquisition loop. Realized label-wise reduction is assigned back only after delayed labels arrive and updates the feature-value model.

5 Evaluation Hypotheses

H1 Under heterogeneous label-local drift, DriftBudget should shift acquisition mass toward features with positive realized risk reduction for the perturbed labels and reduce post-drift class-balanced risk at matched cost.

H2 Under stationary data or a purely global shift, a global-drift acquisition baseline should be equivalent within a pre-specified margin.

H3 Increasing feedback delay should postpone posterior localization and recovery, and the localization error should predict the loss of acquisition value.

H4 Under conditional-feature misspecification or absent support, predicted feature value should cease to track realized reduction and no advantage should be claimed.

Table 1: Claim–evidence alignment used in the evaluation.

Claim	Required experiment	Primary endpoint	Falsifying observation
Label-local allocation	Heterogeneous/global drift factorial	Matched-cost risk and acquisition shift	No method-by-heterogeneity interaction
Risk-value mechanism	Counterfactual feature intervention	Value–reduction rank correlation	Correlation near zero or negative
Delay behavior	Controlled delay sweep	Localization and recovery delay	Diagnostic does not track delay
Boundary of use	Misspecification and no-support stress	Calibration and regret	Advantage persists despite invalid value estimates

Table 2: Results on the chronological MultiEURLEX stream. Cost is normalized to the full-feature policy and recovery is measured in stream steps.

Method	Macro-F1 ↑	Tail recall ↑	Cost ↓	Recovery ↓
Fixed-cheap	59.4	40.8	34.7	28.6
EDDI	64.2	46.9	48.1	22.4
DiFA	66.1	50.2	46.5	17.9
Online AFA tree	66.7	51.1	45.8	16.6
Drift-aware all features	67.0	53.1	100.0	15.7
DRIFTBUDGET w/o delayed credit	69.1	56.4	44.2	13.2
DRIFTBUDGET	**70.4**	**58.0**	44.5	**10.8**
Oracle acquisition	72.2	60.3	44.0	8.5

6 EXPERIMENTAL PROTOCOL

P0 uses a controllable multi-label stream with block-sparse label–feature dependence. The factorial design crosses heterogeneous versus global drift, median delay in $\{0, 2, 5, 10, 20\}$, head versus tail labels, and five budget levels. Monte Carlo intervention supplies ground-truth conditional risk reduction, so the internal value estimate is tested directly rather than inferred from final F1.

The broader P1 protocol specifies chronological MultiEURLEX and MIMIC-III evaluations (Chalkidis et al., 2021; Johnson et al., 2016). Optional-feature groups, costs, delay mechanisms, splits, predictor capacity, conditional-feature model, optimization budget, and stopping rule are fixed before training. Required baselines are fixed cheap features, EDDI, DiFA, explainability-driven AFA, online active-feature trees, static cost-aware acquisition, global-drift DiFA, a full-feature reference, an oracle drift-boundary detector, and an oracle acquisition policy. Retrospective estimators are reported only when their identification assumptions pass the registered diagnostics. The quantitative real-stream result presented in this manuscript is the chronological MultiEURLEX comparison in Table 2; no cross-dataset generalization claim is made from that table.

Primary endpoints are class-balanced risk and tail recall at matched cost, area under the risk–cost curve, posterior-localization delay, recovery delay, acquisition-mass shift, and rank correlation between predicted and realized feature value. At least five seeds are used. The killer intervention permutes label–feature dependence while preserving marginal uncertainty: uncertainty-driven acquisition may continue to purchase a feature, whereas a valid risk-value estimate should collapse when realized reduction vanishes.

7 RESULTS

Table 2 shows that DRIFTBUDGET improves macro-F1 by 4.3 points over DiFA at comparable normalized cost and reduces recovery time from 17.9 to 10.8 steps. The similar cost constrains a resource-volume explanation, although it does not by itself identify the acquisition signal. The full method also exceeds the "w/o delayed credit" variant while using nearly the same cost, separating temporal credit assignment from simply purchasing more features. Removing the label-wise posterior lowers macro-F1 to 66.8 and extends recovery to 18.5 steps; replacing it with a global drift score

yields 68.1 and 15.8. Conditional on the shared implementation, these controls make label locality the main observed distinction from generic adaptivity, with delayed credit providing an additional contribution. In the delay sweep, DRIFTBUDGET retains 65.7 macro-F1 at a median delay of 20 steps, compared with 59.9 for DiFA. This comparison supports task-level robustness to delayed feedback, but the unreported value-ranking and localization endpoints are still needed to attribute that behavior directly to Eq. 4.

8 SCOPE, IMPLICATIONS, AND FAILURE MODES

The reported comparisons support a coherent evidence chain: comparable cost constrains the resource explanation, the global-state and posterior ablations probe label locality, and delayed-credit removal probes temporal assignment. They do not yet show every endpoint registered in Table 1. In particular, the main results do not report the stationary/global equivalence margin, the predicted-versus-realized value rank correlation, or the misspecification and no-support stress tests. Those missing analyses limit how strongly the performance difference can be attributed to Eq. 4 rather than to correlated implementation choices.

The label-local state is not identifiable when feedback is too sparse or when predictor updates mimic distributional change. Risk reduction depends on a conditional feature model that may itself fail under shift, and greedy acquisition is not optimal for arbitrary feature interactions. The evidence is currently limited to the reported MultiEURLEX stream. Accordingly, the mechanism claim applies when posterior localization, value ranking, and heterogeneous-drift benefit agree; it does not establish general superiority under stationary, purely global, or unsupported shifts.

REPRODUCIBILITY STATEMENT

This manuscript specifies the formulation, assumptions, evaluation protocol, and aggregate results reported here. The current anonymous submission package contains the manuscript source and compile dependencies, but not code, experiment configurations, or raw logs; implementation-level reproduction is therefore outside the scope of this package.

ETHICS STATEMENT

The benchmarks must be used under applicable licenses and data-use agreements. Cost-sensitive acquisition can allocate measurements unevenly across labels and populations; aggregate efficiency should not replace audits of tail-label errors and access burdens. These benchmark results do not justify autonomous clinical or legal decisions.

9 CONCLUSION

Multi-label acquisition after drift should purchase measurements for the risks that changed, not for the stream in aggregate. Drift-conditioned value of information couples a delayed label-local drift posterior to feature-specific decision-risk reduction. On the reported stream, this coupling improves macro-F1 and tail recall at lower normalized cost and shortens recovery, while the global-state and delay controls separate the label-local mechanism from generic adaptive acquisition. The broader implication is a decision rule for delayed streams: drift evidence matters operationally only when it changes the ranking of available measurements.

AI USE STATEMENT

AI-assisted tools were used to restructure and render the method illustrations and to update their accompanying descriptions. The scientific content represented in the figures was taken from the manuscript.

REFERENCES

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ilias Chalkidis and Anders Søgaard. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2441–2454, 2022. doi: 10.18653/v1/2022.findings-acl.192.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6974–6996, 2021. doi: 10.18653/v1/2021.emnlp-main.559.

Olivier Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1097–1105, 2014.

Aritra Ghosh and Andrew S. Lan. Difa: Differentiable feature acquisition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7705–7713, 2023. doi: 10.1609/aaai.v37i6.25934.

Osman Berke Guney, Ketan Suhaas Saichandran, Karim Elzokm, Ziming Zhang, and Vijaya B. Kolachalama. Active feature acquisition via explainability-driven ranking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20748–20765, 2025.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1453–1461, 2013.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4234–4243, 2019.

Arman Rahbar, Ziyu Ye, Yuxin Chen, and Morteza Haghir Chehreghani. Efficient online decision tree learning with active feature acquisition. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4163–4171, 2023. doi: 10.24963/ijcai.2023/463.

Santosh T.y.s.s, Tuan-Quang Vuong, and Matthias Grabmair. Chronoslex: Time-aware incremental training for temporal generalization of legal classification tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3022–3039, 2024. doi: 10.18653/v1/2024.acl-long.166.

Henrik von Kleist, Alireza Zamanian, Ilya Shpitser, and Narges Ahmidi. Evaluation of active feature acquisition methods for time-varying feature settings. *Journal of Machine Learning Research*, 26 (60):1–84, 2025.