MEDT: DELAYED-FEEDBACK EXPERT SELECTION FOR STREAMING FRAUD DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming fraud detection must adapt to evolving transaction behavior while recent labels may remain unavailable under delayed supervision. Earlier fraud characteristics may persist, coexist, or reappear, creating the problem of deciding which retained historical models should participate before the newest supervision arrives. We propose **Mixed Expert Decision Trees (MEDT)**, which retains temporally differentiated experts, reassesses predictive competence from observable delayed feedback, and selects high-scoring eligible experts for committee prediction. A frozen historical anchor provides a standing reference, and committee probabilities are combined through anchor-stabilized probability fusion. On the IEEE-CIS-based W26 controlled-perturbation stream, MEDT achieves a PR-AUC of 0.4628, 17.0% above the strongest evaluated online baseline. Removing the temporal-expert component reduces PR-AUC by 0.0436, while replacing soft fusion with hard voting reduces it by 0.1023; both contrasts have Holm-corrected $p < 0.05$ in the planned within-stream paired tests. These results support the temporal-expert component as implemented and soft probability fusion, without establishing a uniquely optimal competence or selection rule.

1 INTRODUCTION

Online payment fraud evolves adversarially as fraudsters adjust strategies, tools, and behavioral patterns to evade existing detection mechanisms. As these behaviors change, the statistical relationships supporting fraud prediction may change, creating a practical need for timely adaptation. A model learned only from earlier observations may therefore become less representative of emerging behavior even while some of its predictive knowledge remains relevant.

Timely adaptation is constrained by delayed supervision. Reliable fraud labels often become available only after investigation, verification, or chargeback processes (Dal Pozzolo et al., 2015; 2018). Consequently, current predictions can use supervised evidence only from earlier label-visible periods, so the newest supervised evidence necessarily lags the transaction stream.

Fraud evolution is also not necessarily a simple replacement process. Previously observed fraud characteristics may persist, coexist with emerging behavior, or reappear later. Different fraudsters may use different strategies at the same time, and earlier tools or behavioral patterns may continue to be reused. Models learned in earlier periods may therefore retain predictive relevance even after newer models are introduced. Adaptation is thus broader than continually replacing an earlier predictive state with the newest one.

These conditions create a dual learning requirement: adapt to emerging fraud behavior while retaining access to historical predictive knowledge that may remain useful. Retention alone, however, does not determine whether an earlier model should participate in current prediction. Because labels describing the newest stream conditions remain unavailable, the predictive competence of retained models for current use cannot be observed directly and must be reassessed from feedback that has already become observable. The research problem is how to use delayed observable feedback to reassess models learned from different historical periods and determine which should participate now.

This problem is narrower than generic adaptation to changing data distributions. Concept-drift adaptation addresses changes in predictive conditions, while historical-model retention keeps ear-

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.

lier knowledge available; neither framing alone specifies which retained models should participate when the newest supervision is absent. We study their narrower intersection: repeated predictive-competence reassessment and explicit committee selection over temporally differentiated experts using only delayed observable feedback. Delayed supervision, historical-model retention, expert selection, and probability fusion are not treated as novel in isolation.

Accordingly, we focus on deciding current model participation under delayed observable feedback rather than detecting recurrence.

We propose **Mixed Expert Decision Trees (MEDT)** for this problem. MEDT retains temporally differentiated experts learned from different stream periods, reassesses their competence as delayed labels become observable, and selects high-scoring eligible experts for a prediction committee. A frozen historical anchor provides a standing reference, and the committee produces fraud-probability predictions through soft fusion. On the IEEE-CIS-based W26 controlled-perturbation stream, MEDT achieves PR-AUC 0.4628, a 17.0% improvement over the strongest evaluated online baseline.

2 Related Work

2.1 Streaming Adaptation under Delayed Fraud Supervision

Streaming fraud detection requires adaptation as predictive conditions evolve. General concept-drift work organizes responses to changing distributions (Gama et al., 2014; Lu et al., 2019). VFDT and EFDT incrementally revise decision trees, iGBDT updates boosting models from newly labeled batches, and HAT/ARF add drift-aware structural or ensemble adaptation (Domingos & Hulten, 2000; Manapragada et al., 2018; Zhang et al., 2019; Bifet & Gavaldà, 2009; Gomes et al., 2017). Delayed supervision introduces a separate timing constraint: feedback for earlier decisions may become observable only later (Joulani et al., 2013). In fraud detection, Dal Pozzolo et al. distinguish recent investigator feedback from later verified labels (Dal Pozzolo et al., 2015); Paldino et al. evaluate historical ensembles under verification latency (Paldino et al., 2024); and continual fraud studies assess delayed feedback through the stability–plasticity lens (Lebichot et al., 2024). ARF also evaluates delayed labeling by postponing label-dependent updates (Gomes et al., 2017). Thus, delayed supervision is established; it does not by itself determine how earlier predictive models should participate when newer stream conditions emerge.

2.2 Historical Predictive Knowledge under Persistent, Coexisting, and Recurring Patterns

Historical-model retention and reuse are likewise established. Learn++.NSE retains classifiers from earlier batches and repeatedly reweights them on newly labeled batches, allowing earlier classifiers to regain influence under cyclical change (Elwell & Polikar, 2011). Paldino et al. evaluate historical fraud-model ensembles under verification latency using diversity- and performance-based mechanisms (Paldino et al., 2024). Dal Pozzolo et al. retain a bounded delayed ensemble, while Lebichot et al. study preservation and forgetting of earlier knowledge (Dal Pozzolo et al., 2015; Lebichot et al., 2024). These works show that retention itself is not new. Retaining a model differs from deciding whether it should participate in the next prediction period when the newest stream labels are unavailable.

2.3 Delayed Competence Reassessment and Expert Committee Formation

Adaptive ensembles determine participation through weighting, pruning, or replacement. DWM changes expert weights from current errors and removes weak experts; AUE2 reweights and replaces chunk models; Learn++.NSE repeatedly reweights retained classifiers (Kolter & Maloof, 2007; Brzeziński & Stefanowski, 2014; Elwell & Polikar, 2011). Dynamic selection explicitly estimates classifier competence to select predictors for a given decision (Cruz et al., 2018). MEDT instead forms a segment-level committee that is reconsidered when new delayed feedback becomes observable. Paldino et al. evaluate historical fraud-model ensembles under verification latency using diversity- and performance-based mechanisms (Paldino et al., 2024). MEDT studies the narrower intersection of these mechanisms: retained temporal experts are repeatedly reassessed using only

delayed label-visible feedback, and high-scoring experts are selected for the next segment. This positioning concerns delayed predictive-competence reassessment and temporal-expert selection, not concept-drift or explicit recurrence detection.

3 METHODOLOGY

MEDT follows five steps: *retain*, *reassess*, *select*, *form a committee*, and *predict*. Temporally differentiated experts provide historical candidates; delayed observable feedback reassesses their current-use competence; selection forms an active subset; and a frozen historical anchor joins the committee for anchor-stabilized probability fusion. The estimator, selection rule, and anchor are design choices, not claims of independent optimality or gain.

3.1 FRAMEWORK OVERVIEW

Let the transaction stream be partitioned into ordered temporal segments $\{\mathcal{D}_t\}_{t=0}^{T}$, and let L denote the label delay. At prediction time t, only segments

$$\mathcal{V}_t = \{u \mid u \leq t - L\} \tag{1}$$

are label-visible. This visibility boundary constrains all label-dependent operations: a temporal expert can be trained only after its segment becomes label-visible, and its competence can be reassessed only from later feedback that has also crossed the boundary.

The boundary separates predictive knowledge generation from competence estimation: frozen experts preserve historical candidates, while delayed observable feedback determines current participation. Selection is distinct from retention, and fusion converts the selected committee into a transaction-level probability.

The anchor M_{base} is trained once on the fully labeled warm-up training and validation data, remains frozen, and joins every committee. A post-warm-up segment yields a temporal expert when label-visible; that expert becomes eligible only after later competence feedback. MEDT then selects up to N high-scoring eligible experts for the current segment.

3.2 TEMPORALLY DIFFERENTIATED EXPERT RETENTION

Earlier predictive relationships may remain relevant even after newer behavior emerges. MEDT therefore keeps experts learned from different historical periods available instead of treating the newest model as the only candidate predictor. Retention provides availability; it does not assume that an earlier expert remains useful for current prediction.

Once segment w becomes label-visible, MEDT trains

$$M_w = f\left(\mathcal{D}_w^{\text{train}} \cup \mathcal{D}_w^{\text{val}}\right), \tag{2}$$

where f denotes the common gradient-boosted decision-tree learning procedure. The resulting expert M_w is frozen for subsequent competence reassessment and selection. Let W_{start} denote the first post-warm-up segment. The temporal experts available at time t are

$$\mathcal{C}_t = \{M_w \mid w \in \mathcal{V}_t,\ w \geq W_{\text{start}}\}. \tag{3}$$

A temporal segment is an operational stream unit rather than a ground-truth fraud mode. Thus, temporal differentiation describes when an expert was trained, not specialization to a uniquely identifiable fraud pattern.

3.3 DELAYED-FEEDBACK COMPETENCE REASSESSMENT

Retained experts can differ in their current usefulness, but the newest transaction outcomes are unavailable under delayed supervision. MEDT therefore uses only later label-visible validation feedback to construct a competence score for current participation. Here, *competence* denotes predictive ability relevant to committee participation; the implemented score is an observable proxy rather than direct knowledge of true current reliability.

MEDT uses a precision-oriented signal to emphasize false-positive control in the sparse-positive fraud setting. This is an operational design choice, not a claim that precision is the uniquely correct definition of competence. For temporal expert M_w evaluated on the validation subset of a later label-visible segment u, let $\mathrm{TP}_u^{(w)}$ and $\mathrm{FP}_u^{(w)}$ denote counts under the fixed implementation threshold of 0.5, which is not assumed to be universally optimal. Define

$$n_u^{(w)} = \mathrm{TP}_u^{(w)} + \mathrm{FP}_u^{(w)}, \qquad \hat{p}_u^{(w)} = \frac{\mathrm{TP}_u^{(w)}}{n_u^{(w)}}. \tag{4}$$

Because raw precision can be unstable when the number of positive predictions is small, MEDT uses the Wilson lower bound:

$$\mathrm{LB}_u^{(w)} = \frac{\hat{p}_u^{(w)} + \frac{z^2}{2n_u^{(w)}} - z\sqrt{\frac{\hat{p}_u^{(w)}(1-\hat{p}_u^{(w)})}{n_u^{(w)}} + \frac{z^2}{4(n_u^{(w)})^2}}}{1 + \frac{z^2}{n_u^{(w)}}}, \tag{5}$$

for $n_u^{(w)} > 0$, with $z = 1.96$ in the experiments.

Only later label-visible segments provide competence evidence for M_w:

$$\mathcal{U}_w(t) = \left\{ u \in \mathcal{V}_t \mid u > w,\ n_u^{(w)} > 0 \right\}. \tag{6}$$

The condition $u > w$ prevents assessment on the expert's own training segment, while $n_u^{(w)} > 0$ excludes undefined precision observations. If a later segment u was observed before M_w was trained, the frozen expert is replayed on its buffered validation features. These replayed scores are used only for subsequent competence reassessment and do not alter predictions already issued; the corresponding labels enter the reassessment only after they become visible.

As delayed evidence accumulates across later segments, MEDT aggregates the available Wilson observations with recency weighting rather than relying on a single evaluation segment. Let $v_t = \max(\mathcal{V}_t)$ be the most recent label-visible segment. MEDT computes

$$\mathrm{Score}_w(t) = \frac{\sum_{u \in \mathcal{U}_w(t)} \gamma^{v_t - u} \mathrm{LB}_u^{(w)}}{\sum_{u \in \mathcal{U}_w(t)} \gamma^{v_t - u}}, \tag{7}$$

where $\gamma \in (0, 1]$ gives more weight to recent label-visible feedback when $\gamma < 1$. Repeated reassessment is a mechanism for incorporating successively observable delayed evidence. It is not claimed or proven to be inherently more stable than one-shot assessment.

3.4 Competence-Based Expert Selection

Retention maximizes the availability of historical predictive knowledge, but availability should not automatically imply current participation. MEDT therefore converts the retained pool into a bounded active subset using only competence evidence that is already observable.

Only experts with later competence feedback are eligible:

$$\mathcal{C}_t^{\mathrm{elig}} = \{ M_w \in \mathcal{C}_t \mid \mathcal{U}_w(t) \neq \emptyset \}. \tag{8}$$

MEDT then selects up to N highest-scoring eligible experts,

$$\mathcal{E}_t = \mathrm{TopN}_{M_w \in \mathcal{C}_t^{\mathrm{elig}}} \left(\mathrm{Score}_w(t) \right), \tag{9}$$

using all eligible experts when fewer than N are available. This step operationalizes the distinction between historical availability and current committee participation. Wilson–EWMA is the default competence estimator and Top-N ranking is the default selection rule; neither is presented as uniquely optimal or as an independently established source of MEDT's predictive gain.

3.5 Anchor-Stabilized Probability Fusion

Selection determines which temporal experts participate, but the committee still requires a single fraud-probability output. The selected temporal subset changes as delayed evidence becomes observable and can also be empty before any temporal expert becomes eligible. MEDT therefore

retains the frozen historical anchor as a standing committee member and combines member probabilities directly.

The selected temporal experts and the anchor form

$$\mathcal{K}_t = \mathcal{E}_t \cup \{M_{\text{base}}\}. \tag{10}$$

The anchor is trained once from the fully labeled warm-up history, spans a broader historical period than any single temporal expert, remains frozen in the present implementation, and always participates in the final combination. We use *anchor-stabilized probability fusion* as a compact design label: the anchor provides a persistent historical reference and keeps the prediction path well-defined when no temporal expert is eligible. The term does not imply an independently demonstrated improvement in predictive stability.

MEDT combines fraud-probability estimates rather than first reducing each model output to a binary vote, preserving probability magnitude for committee prediction. For transaction x_i,

$$\tilde{p}_i = \frac{w_{\text{base}} p_i^{\text{base}} + \sum_{j \in \mathcal{E}_t} p_i^{(j)}}{w_{\text{base}} + |\mathcal{E}_t|}, \tag{11}$$

with

$$w_{\text{base}} = \max\left(1, \left\lfloor \frac{N+1}{2} \right\rfloor\right). \tag{12}$$

This fixed heuristic assigns the historical anchor a weight of roughly half the target temporal-committee size, preserving a persistent warm-up contribution without giving the anchor the full weight of the temporal committee; it is not tuned on test performance. When no temporal expert is eligible, the committee prediction reduces to the anchor prediction. Its role is operational: to maintain a persistent reference and produce one committee probability; any independent predictive or stability benefit remains an empirical question.

The complete update order is given in Algorithm 1 in the Appendix.

4 Experimental Setup

The experiments address four method-aligned questions: overall effectiveness under chronological delayed supervision (Q1); contribution of the temporal-expert component and later selection of earlier experts (Q2); robustness to competence and selection choices (Q3); and the roles of soft fusion and the historical anchor in committee prediction (Q4). Bundled ablations are interpreted only at the component level, while non-significant contrasts do not establish independent effects.

4.1 Chronological Delayed-Supervision Evaluation

We construct the 26-week W26 stream from the IEEE-CIS Fraud Detection dataset (Vesta Corporation & IEEE-CIS, 2019), which contains 590,540 transactions and 20,663 fraud cases (3.5%). Transactions are ordered chronologically into weekly segments. The default evaluation uses label delay $L = 2$, five warm-up weeks, and 19 prediction weeks. Within each week, the first 55% of transactions are used for training, the next 20% for validation, and the final 25% for held-out testing, preserving order without shuffling.

At prediction time t, only labels from $u \leq t - L$ are observable. Post-warm-up expert training, competence reassessment, and selection respect this boundary. Once a segment becomes label-visible, its training and validation portions are used to train and freeze the new temporal expert; validation feedback also reassesses previously retained experts from stored predictions. The new expert is not reassessed on its own segment, and test labels are evaluation-only.

4.2 Baseline Roles and Fair Comparison

VFDT and HAT provide classical and drift-adaptive tree references; ARF is a drift-adaptive online ensemble. Streaming HGB and iGBDT provide streaming/incremental boosting comparisons. DWM, Learn++.NSE, and AUE2 represent adaptive ensembles based on weighting, block updates,

or replacement, with Learn++.NSE the clearest evaluated historical-retention comparator. XGBoost (Chen & Guestrin, 2016) is a batch reference refit on all label-visible training and validation history.

All methods use the same chronological partitions, shared input representation after common preprocessing, delayed label-visibility schedule, and held-out test transactions. Model-specific transformations and updates are restricted to information available at the same prediction time. Baseline update schemes are reported in Appendix B, with complete configurations in the supplementary artifact.

4.3 Metrics and Statistical Comparisons

PR-AUC is primary because the stream is imbalanced. $F1^{\tau^*}$, Recall, and ECE provide complementary threshold-dependent detection and calibration views. For each method and prediction week, τ^* is selected retrospectively by maximizing F1 on that week's validation labels and then applied unchanged to test scores. It is not used for model training, competence reassessment, or selection; $F1^{\tau^*}$ and Recall are therefore retrospective diagnostics rather than deployable same-week operating performance. Competence reassessment and hard voting use the fixed threshold 0.5.

For the six pre-specified primary online baselines, paired weekly PR-AUC and $F1^{\tau^*}$ comparisons use one-sided Wilcoxon signed-rank tests with Holm correction. VFDT/HAT remain additional classical references and XGBoost a batch reference outside this inferential family. Component contrasts use planned one- or two-sided Wilcoxon tests according to their hypotheses, again with Holm correction. Because all 19 paired observations come from one chronological stream, serial dependence may remain; these tests are interpreted only as within-stream paired evidence rather than independent replication. Full test specifications and adjusted p-values are in Appendix C.

4.4 Component Ablations and Parameter Sensitivity

The `w/o temporal experts` variant removes the temporal-expert component jointly, including later reassessment, selection, and temporal contributions to fusion. Competence-rule variants alter the score; uniform recent-N replaces score-based selection; and `w/o base expert` removes the anchor while retaining soft fusion. Hard voting keeps the selected experts and anchor weight but replaces probabilities with binary votes at 0.5, providing a more isolated contrast for soft fusion.

Sensitivity varies one factor at a time over label delay L, committee size N, and EWMA decay γ, with defaults $L = 2$, $N = 5$, and $\gamma = 0.85$. Nine non-default configurations are compared with the default using paired two-sided Wilcoxon tests with Holm correction. Complete settings and results are reported in Appendix C.3; the main text retains only their aggregate interpretation.

4.5 Controlled Temporal Stress and Mechanism Diagnostics

W26 is an evaluation/stress construction, not a MEDT methodological contribution. Its mixed schedule applies Amount Mimicry, Credential Churn, Modus Operandi Shift, and Traffic Reordering to fraud transactions; legitimate rows, labels, and sample counts remain unchanged. The main 19-week evaluation preserves the delayed-supervision protocol. Isolated single-perturbation diagnostics in Appendix D characterize post-change performance, later selection of earlier experts, and W26 protocol sensitivity. They can establish continued selection of earlier experts, but not their individual causal contribution, recurrence detection, or strategic attacker behavior.

5 Results and Discussion

5.1 Q1: Overall Effectiveness under Delayed Supervision

Table 1 answers Q1 on the IEEE-CIS-based W26 chronological delayed-supervision stream. MEDT achieves macro PR-AUC 0.4628, the highest among evaluated online methods. Relative to iGBDT (0.3955), the strongest online PR-AUC baseline, this is a 17.0% relative improvement. MEDT also has the highest retrospective $F1^{\tau^*}$ and Recall among online methods. Paired weekly PR-AUC and $F1^{\tau^*}$ comparisons against all six pre-specified primary online baselines yield Holm-corrected $p < 0.05$ in the planned within-stream tests; full statistics are in Appendix C.

Table 1: Overall performance on the IEEE-CIS-based W26 delayed-supervision stream, macro-averaged over 19 evaluation weeks. XGBoost is included as a batch retraining reference under the same label-visibility constraint. Bold denotes the best online result for each metric (lower is better for ECE).

Method	**PR-AUC**	**F1$^{\tau^*}$**	**Recall**	**ECE**
XGBoost (retraining ref.)	0.4829	0.4709	0.4049	0.2448
VFDT	0.2223	0.3177	0.2686	0.0161
HAT	0.1796	0.2699	0.2490	0.0236
ARF	0.3253	0.3369	0.2839	**0.0127**
Streaming HGB	0.3287	0.3774	0.3224	0.1934
iGBDT	0.3955	0.4059	0.3322	0.3926
DWM	0.3520	0.3639	0.3303	0.1763
Learn++.NSE	0.3461	0.3616	0.2919	0.1644
AUE2	0.3339	0.3635	0.3240	0.2267
MEDT	**0.4628**	**0.4546**	**0.3768**	0.1071

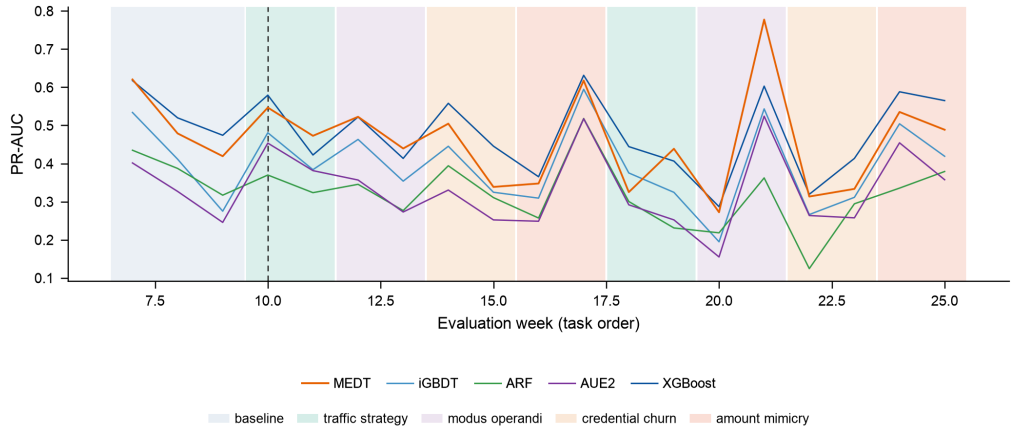


Figure 1: Weekly PR-AUC trajectories for MEDT and representative comparators over the 19 delayed-supervision prediction weeks (weeks 7–25) under the default W26 mixed-perturbation schedule. Shaded bands indicate the perturbation modes. XGBoost is shown as a batch retraining reference under the same label-visibility constraint.

XGBoost attains higher aggregate PR-AUC, $F1^{\tau^*}$, and Recall than MEDT, but it is a batch reference refit on all label-visible history rather than a directly comparable online learner. MEDT also does not dominate calibration: ARF has the lowest ECE. Thus the Q1 result is specific to detection performance on the evaluated W26 delayed-supervision stream and should not be read as universal superiority. Figure 1 shows the corresponding weekly PR-AUC trajectories.

5.2 Q2: Temporal-Expert Contribution

Removing the entire temporal-expert component reduces PR-AUC from 0.4628 to 0.4192 ($\Delta\mathrm{PR} = +0.0436$; $p_{\mathrm{Holm}} = 0.0029$). This supports the component as implemented. Because the variant also removes later reassessment, selection, and temporal contributions to fusion, the effect cannot be attributed to retention alone.

W26 diagnostics provide complementary mechanism-level evidence. MEDT's final post-change/pre-change PR-AUC ratio ranges from 78.2% to 81.0% across isolated perturbations; the ratio is descriptive and not a stability or recovery guarantee. MEDT's absolute final post-change performance is higher than ARF and AUE2 after all four perturbations, while the comparison with iGBDT is mixed across perturbation types. In the mixed main evaluation, the expert trained on week 10 remains among the Top-5 at week 25, showing that earlier experts can remain selected

Table 2: Component analysis of MEDT over 19 evaluation weeks. ΔPR denotes Full MEDT minus the corresponding variant. Holm correction is applied over the six planned contrasts.

Setting	PR-AUC	$F1^{\tau^*}$	Recall	ΔPR	p_{Holm}
MEDT (full)	0.4628	0.4546	0.3768	—	—
Hard voting	0.3605	0.4262	0.3846	+0.1023	1.1×10^{-5}
w/o temporal experts	0.4192	0.4236	0.3420	+0.0436	**0.0029**
Raw precision	0.4782	0.4669	0.4010	-0.0154	0.4667
Wilson–EWMA (accuracy)	0.4662	0.4576	0.3916	-0.0035	1.0000
Uniform recent-N	0.4633	0.4507	0.3704	-0.0005	1.0000
w/o base expert	0.4727	0.4658	0.4063	-0.0099	1.0000

later. This establishes continued selection, not individual causal contribution or recurrence detection (Appendix D).

5.3 Q3: Competence and Selection Robustness

The competence and selection variants show no independent advantage for the defaults. Raw precision reaches PR-AUC 0.4782 and does not yield a significant difference from Full MEDT under the planned two-sided comparison. Accuracy-oriented Wilson–EWMA reaches 0.4662, and uniform recent-N selection reaches 0.4633; neither yields a significant degradation after correction. The evidence therefore does not attribute MEDT's overall gain to the specific precision-oriented estimator or Top-N rule.

Across the tested settings, PR-AUC ranges from 0.4589 to 0.4649 for L, from 0.4597 to 0.4628 for N, and changes by less than 0.0005 around the default for γ; none of the nine planned contrasts is significant after Holm correction. These non-significant results do not establish statistical equivalence or imply that delay is irrelevant. The sensitivity table and paired tests are in Appendix C.3.

5.4 Q4: Committee Prediction

Replacing soft probability fusion with hard voting produces the largest degradation, reducing PR-AUC from 0.4628 to 0.3605 ($\Delta\mathrm{PR} = +0.1023$; $p_{\mathrm{Holm}} = 1.1 \times 10^{-5}$). Because committee membership and anchor weight are unchanged, this directly supports soft fusion relative to the implemented hard-voting alternative. Removing the anchor yields PR-AUC 0.4727 with no significant two-sided difference, so its independent predictive or stability benefit remains inconclusive.

5.5 Discussion and Limitations

Together, Q1–Q4 support an architecture-level interpretation rather than independent gains from every MEDT design choice. MEDT achieves the highest macro PR-AUC among evaluated online methods on the W26 delayed-supervision stream; the temporal-expert component contributes as implemented; and soft fusion has the clearest isolated component evidence. In contrast, the experiments do not identify a uniquely necessary competence estimator or selection rule, and the ablations do not isolate competence-based selection itself as the causal source of the overall gain. The anchor remains a persistent historical reference by design rather than an independently validated performance source.

Several limitations bound these conclusions. The competence signal is delayed by L and cannot observe changes inside the current label-blind window. The normalized score is updated only when a valid precision observation exists, and the fixed 0.5 competence threshold may depend on calibration. The experiments do not compare repeated reassessment with a one-shot alternative, isolate individual historical-expert contributions, or establish fraud-mode specialization. Finally, W26 is constructed from the single public IEEE-CIS dataset and uses controlled perturbations rather than strategic attacker behavior or explicit recurrence detection. Extended sensitivity, computational measurements, and controlled temporal diagnostics are reported in the Appendix.

6 CONCLUSION

We study streaming transaction fraud detection under delayed supervision, where recent labels may be unavailable while earlier and emerging behaviors may coexist or reappear. MEDT retains temporally differentiated experts, reassesses their predictive competence from later label-visible feedback, selects high-scoring eligible experts for each prediction segment, and combines them with a frozen historical anchor through anchor-stabilized probability fusion.

On the IEEE-CIS-based W26 controlled-perturbation stream, MEDT achieves PR-AUC 0.4628, 17.0% above the strongest evaluated online baseline. Removing the temporal-expert component and replacing soft fusion with hard voting both yield Holm-corrected $p < 0.05$ in the planned within-stream paired tests. The evidence therefore supports the temporal-expert component as implemented and soft fusion relative to the tested hard-voting alternative. In contrast, competence, selection, and anchor contrasts are not significant after correction, so they do not establish a uniquely optimal estimator, selection rule, or independent anchor benefit. Selection traces show that earlier experts can remain selected later, but not their individual causal contribution or recurrence detection. The evaluation remains limited to a single public dataset and controlled W26 perturbations.

ACKNOWLEDGMENTS

We thank Vesta and the organizers of the IEEE-CIS Fraud Detection Challenge for publicly releasing the IEEE-CIS fraud dataset (Vesta Corporation & IEEE-CIS, 2019), which provides the empirical basis for this study.

AI USE STATEMENT

AI tools were used to assist with literature discovery and positioning, feedback on methodology and experimental presentation, interpretation and testing of reported results, supportive coding assistance, translation, and manuscript editing for clarity, structure, and consistency. All AI-assisted suggestions were reviewed by the authors. The authors take responsibility for the research decisions, reported results, citations, and final content of the submission.

ETHICS STATEMENT

This study uses the publicly released and anonymized IEEE-CIS Fraud Detection dataset and involves no direct interaction with human subjects or attempts to re-identify individuals. The work evaluates machine-learning methods for fraud detection in an offline research setting. Because practical fraud detection can affect legitimate transactions and impose costs through false positives, the reported models should not be interpreted as ready for autonomous deployment or as substitutes for appropriate operational review. The controlled perturbations are used solely for methodological evaluation and are not intended to reproduce or facilitate real fraudulent activity.

REFERENCES

Albert Bifet and Ricard Gavaldà. Adaptive learning from evolving data streams. In *Advances in Intelligent Data Analysis VIII*, volume 5772 of *Lecture Notes in Computer Science*, pp. 249–260. Springer, 2009. doi: 10.1007/978-3-642-03915-7_22.

Dariusz Brzeziński and Jerzy Stefanowski. Reacting to different types of concept drift: The accuracy updated ensemble algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, 25 (1):81–94, 2014. doi: 10.1109/TNNLS.2013.2251352.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Rafael M. O. Cruz, Robert Sabourin, and George D. C. Cavalcanti. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216, 2018. doi: 10.1016/j.inffus.2017.09.010.

Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi. Credit card fraud detection and concept-drift adaptation with delayed supervised information. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2015. doi: 10.1109/IJCNN.2015.7280527.

Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi. Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3784–3797, 2018. doi: 10.1109/TNNLS.2017.2736643.

Pedro Domingos and Geoff Hulten. Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 71–80. ACM, 2000. doi: 10.1145/347090.347107.

Ryan Elwell and Robi Polikar. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks*, 22(10):1517–1531, 2011. doi: 10.1109/TNN.2011.2160459.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014. doi: 10.1145/2523813.

Heitor M. Gomes, Albert Bifet, Jesse Read, Jean Paul Barddal, Fabrício Enembreck, Bernhard Pfahringer, Geoff Holmes, and Talel Abdessalem. Adaptive random forests for evolving data stream classification. *Machine Learning*, 106(9–10):1469–1495, 2017. doi: 10.1007/s10994-017-5642-8.

Pooria Joulani, András György, and Csaba Szepesvári. 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. PMLR, 2013. URL https://proceedings.mlr.press/v28/joulani13.html.

J. Zico Kolter and Marcus A. Maloof. Dynamic weighted majority: An ensemble method for drifting concepts. *Journal of Machine Learning Research*, 8:2755–2790, 2007. URL https://www.jmlr.org/papers/v8/kolter07a.html.

Bertrand Lebichot, Wissam Siblini, Gian Marco Paldino, Yann-Aël Le Borgne, Frédéric Oblé, and Gianluca Bontempi. Assessment of catastrophic forgetting in continual credit card fraud detection. *Expert Systems with Applications*, 249:123445, 2024. doi: 10.1016/j.eswa.2024.123445.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2019. doi: 10.1109/TKDE.2018.2876857.

Chaitanya Manapragada, Geoffrey I. Webb, and Mahsa Salehi. Extremely fast decision tree. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1953–1962. ACM, 2018. doi: 10.1145/3219819.3220005.

Gian Marco Paldino, Bertrand Lebichot, Yann-Aël Le Borgne, Wissam Siblini, Frédéric Oblé, Giacomo Boracchi, and Gianluca Bontempi. The role of diversity and ensemble learning in credit card fraud detection. *Advances in Data Analysis and Classification*, 18(1):193–217, 2024. doi: 10.1007/s11634-022-00515-5.

Vesta Corporation and IEEE-CIS. IEEE-CIS fraud detection. Kaggle competition dataset provided by Vesta Corporation, 2019. Accessed 2026. https://www.kaggle.com/c/ieee-fraud-detection.

Chongsheng Zhang, Yuan Zhang, Xianjin Shi, George Almpanidis, Gaojuan Fan, and Xiajiong Shen. On incremental learning for gradient boosting decision trees. *Neural Processing Letters*, 50:957–987, 2019. doi: 10.1007/s11063-019-09999-3.

Appendix navigation. Appendix A covers the evaluation protocol and W26 construction; Appendix B, MEDT and baseline implementation; Appendix C, statistical support; Appendix D, temporal diagnostics; and Appendix E, computational details.

A Evaluation Protocol and W26 Construction

This section specifies the chronological delayed-supervision protocol and W26 construction underlying the main evaluation. It supports the experimental setting summarized in Section 4 and separates evaluation construction from MEDT's methodological contribution.

A.1 Chronological Stream and Label Visibility

Table 3: Default IEEE-CIS-based W26 delayed-supervision evaluation protocol and stream construction.

Configuration	Value
Source dataset	IEEE-CIS Fraud Detection
Transactions	590,540
Fraud cases	20,663 (3.5%)
Ordering	`TransactionDT`
Stream construction	W26, 26 weekly segments
Warm-up	Weeks 0–4
First post-warm-up segment	$W_{\text{start}} = 5$
Reported prediction weeks	Weeks 7–25 (19 weeks)
Train/validation/test	Chronological 55%/20%/25%; no shuffling
Label delay L	2
Main W26 schedule	Mixed perturbations; injection begins at week 10; default span = 2
Reporting threshold	Retrospective τ^* for $F1^{\tau^*}$/Recall

At prediction time t, labels are visible only for segments $u \leq t - L$. Under the default $L = 2$, week 5 is the first post-warm-up segment and becomes label-visible at $t = 7$, yielding 19 reported prediction weeks (weeks 7–25). Post-warm-up training, competence reassessment, and selection use only label-visible information, and test labels are used only for evaluation.

When segment w becomes label-visible, its validation labels are used to reassess previously trained experts from stored predictions or predictions replayed by a subsequently trained frozen expert on buffered validation features. Labels are converted to competence TP/FP counts only after becoming visible. The training and validation portions of w are then used to train and freeze M_w; because competence requires later feedback $u > w$, M_w is never assessed on its own segment.

A.2 W26 Controlled Perturbation Construction

W26 modifies fraud transactions only; legitimate rows, class labels, and sample counts are unchanged. Ground-truth fraud labels are used offline only to build the controlled evaluation stream and are not exposed to the learner before the corresponding visibility boundary. Weekly membership is fixed from the original `TransactionDT` before perturbations are materialized, so perturbation construction does not change the learner's delayed label-visibility schedule.

The main mixed schedule uses unperturbed weeks 0–9 followed by a two-week rotation T–M–C–A: Traffic Reordering (10–11), Modus Operandi Shift (12–13), Credential Churn (14–15), and Amount Mimicry (16–17), followed by the same sequence over weeks 18–25. Isolated diagnostics instead apply one perturbation during weeks 10–13, remove it during weeks 14–17, reintroduce it during weeks 18–21, and remove it again during weeks 22–25. These are controlled distributional stress mechanisms, not simulations of strategic attacker responses or evidence of explicit recurrence detection.

Table 4: Operational definitions of the four W26 perturbations. Traffic Reordering corresponds to `traffic_strategy` in the implementation.

Mode	Transformation applied to fraud transactions	Default intensity
Amount Mimicry	Modify only `TransactionAmt`. Let $s = \max(\sigma_{\text{legit}}, \sigma_{\text{fraud}}, 10^{-6})$ and $z \sim \mathcal{N}(0, (0.15s)^2)$. A target amount is $\mu_{\text{legit}} + z$, and the new amount is $(1-\alpha)a + \alpha(\mu_{\text{legit}} + z)$.	Maximum $\alpha = 0.35$; ramp length $= 8$; noise scale $= 0.15$.
Credential Churn	Resample card, address, device, and `id_*` fields from their non-missing pools. Weekly bursts assign a sampled fraud subset shared replacement values on at least two fields; remaining rows receive independent background resampling. Numeric fields receive additional Gaussian jitter.	Maximum $p_{\text{row}} = 0.30$; ramp length $= 6$; burst probability $= 0.32$; burst fraction $[0.25, 0.68]$; background scale 0.42; numeric jitter 0.4σ.
Modus Operandi Shift	Move a sampled fraud subset to a target `ProductCD` line (by default, the dominant and second-dominant fraud product lines define source and target) and resample available card/email/device fields from donor distributions of the target line. Categorical fields use frequency-based sampling and numeric high-cardinality fields use empirical sampling.	Fraud fraction sampled uniformly from $[0.78, 1.0]$; no continuous α.
Traffic Reordering	Sample fraud rows and group them by `card1`. ProductCD is diversified from a mixture of the fraud empirical distribution and a uniform distribution. Groups whose maximum amount exceeds fraud $Q_{0.75}$ or whose total exceeds $1.5Q_{0.75}$ are fragmented while preserving group amount; isolated large amounts are capped at fraud $Q_{0.25}$. Within-group `TransactionDT` is compressed using a minimum gap derived from the 5th percentile of positive legitimate gaps (clipped to 30–86400 seconds). Weekly segmentation still uses original timestamps.	Fraud fraction $[0.65, 0.95]$; ProductCD diversification probability 0.55 with uniform mixture weight 0.45; amount-split noise 0.025.

Table 5: Evaluation metrics and threshold rules for the W26 delayed-supervision evaluation.

Metric / threshold	Implementation and role
PR-AUC	Computed directly from prediction scores using `average_precision_score`; primary detection metric.
$F1^{\tau^*}$ / Recall	For each method and prediction week, τ^* is selected retrospectively by maximizing F1 on that week's validation labels and then applied unchanged to test scores; it is evaluation-only and not available to the learner at prediction time.
ECE	Computed from prediction scores using 15 equal-width bins over $[0, 1]$, with the final bin including the right endpoint.
Competence / hard vote	A separate fixed threshold of 0.5 is used for temporal-expert TP/FP counts and for the hard-voting ablation.

A.3 Metrics and Threshold Implementation

For ECE,

$$\text{ECE} = \sum_{b=1}^{15} \frac{n_b}{n} \left| \text{acc}_b - \text{conf}_b \right|,$$

where n_b is the number of evaluated instances in bin b, acc_b is the empirical fraud rate, and conf_b is the mean predicted fraud probability. Empty bins contribute zero. Test labels are never used to select τ^*.

B MEDT and Baseline Implementation

This section records the execution order and implementation settings needed to reproduce MEDT and its baselines. It complements the mathematical formulation in Section 3 without repeating the method motivation.

B.1 MEDT Update Order

Algorithm 1 MEDT under delayed supervision

Input: stream segments $\mathcal{D}_t$, label delay L, first post-warm-up segment W_{start}, committee size N
Output: streaming predictions

Train the frozen base expert M_{base} on the training and validation portions of the fully labeled warm-up segments
Initialize the temporal expert pool $\mathcal{C} \leftarrow \emptyset$
for each evaluation segment t **do**
  Set the most recent label-visible segment $v_t \leftarrow t - L$
  if $v_t \geq W_{\text{start}}$ **then**
    Observe the labels of segment v_t and compute TP/FP counts from the stored temporal-expert predictions on its validation transactions
    Reassess each temporal expert $M_w \in \mathcal{C}$ with $w < v_t$ using the validation feedback from segment v_t
    Train the new temporal expert $M_{v_t} \leftarrow f(\mathcal{D}_{v_t}^{\text{train}} \cup \mathcal{D}_{v_t}^{\text{val}})$
    Freeze M_{v_t} and update $\mathcal{C} \leftarrow \mathcal{C} \cup \{M_{v_t}\}$
    Generate and store M_{v_t} predictions on buffered validation features of already observed later segments u with $v_t < u < t$
  end if
  Form the eligible expert pool $\mathcal{C}_t^{\text{elig}} \leftarrow \{M_w \in \mathcal{C} \mid \mathcal{U}_w(t) \neq \emptyset\}$
  Compute $\text{Score}_w(t)$ for each $M_w \in \mathcal{C}_t^{\text{elig}}$ using Eq. 7
  Select the top-N temporal experts $\mathcal{E}_t \leftarrow \text{TopN}_{M_w \in \mathcal{C}_t^{\text{elig}}}(\text{Score}_w(t))$
  Form the prediction committee $\mathcal{K}_t \leftarrow \mathcal{E}_t \cup \{M_{\text{base}}\}$
  Produce committee probability predictions through soft fusion using Eq. 11
  Predict the test transactions in segment t
  Store temporal-expert predictions on the validation transactions of segment t for later competence reassessment when their labels become observable
end for

Table 6: Default MEDT implementation parameters.

Parameter	Value
Committee size N	5
Maximum temporal-expert pool size	32
EWMA decay γ	0.85
Wilson parameter z	1.96
Base-expert weight w_{base}	3
Competence / hard-vote threshold	0.5

B.2 MEDT Configuration

The temporal-expert pool is capped at 32 models. This capacity is not reached in the 26-week W26 evaluation, so retirement behavior after saturation is not assessed and does not affect the reported results.

B.3 Baseline Update Schemes

All baselines use the same chronological stream, data partitions, and label-visibility schedule as MEDT. No baseline update uses labels that have not crossed the common visibility boundary, and test labels are excluded from training and updating.

B.4 Preprocessing and Hyperparameter Availability

All methods use the shared input representation after common preprocessing described by the experiment implementation, and complete model-specific hyperparameters and executable configurations are provided in the supplementary artifact.

Table 7: Implemented baseline update schemes under the common W26 delayed-supervision protocol.

Method	Update scheme
XGBoost (retraining ref.)	Weekly batch refit using all label-visible training and validation history; past test labels are excluded
Streaming HGB	Weekly warm-start updates in the implemented `HistGradientBoostingClassifier` configuration
ARF / VFDT / HAT	River incremental implementations; updates occur only when labels become visible
iGBDT	Incremental updates under the same delayed-supervision schedule
DWM / Learn++.NSE / AUE2	Chunk-ensemble updates when each newly label-visible week becomes available

C Statistical Support for Main Results

This section provides the paired statistical evidence behind the main Q1–Q4 claims. The prediction week is the paired observation: all tests use matched metric values from the same 19 prediction weeks. Because these weeks belong to one sequential stream, serial dependence may remain; the resulting tests are interpreted only as within-stream paired evidence rather than independent replication. Reported Δ values are mean paired weekly differences and do not establish cross-stream or cross-dataset generalization.

C.1 Q1: MEDT versus Online Baselines

The primary comparison family contains six planned online baselines: ARF, Streaming HGB, iGBDT, DWM, Learn++.NSE, and AUE2. Paired one-sided Wilcoxon signed-rank tests evaluate $H_1 : \mathrm{MEDT} > \mathrm{baseline}$ over the 19 matched prediction weeks, with Holm correction applied separately to the PR-AUC and $F1^{\tau^*}$ families ($m = 6$). VFDT/HAT remain classical references and XGBoost remains a batch retraining reference outside this inferential family.

Table 8: Paired one-sided Wilcoxon comparisons between MEDT and the six primary online baselines over 19 matched prediction weeks (Holm family $m = 6$). Positive Δ indicates a higher mean weekly metric for MEDT.

Baseline	ΔPR	p_{PR}	$p_{\mathrm{PR}}^{\mathrm{Holm}}$	$\Delta F1^{\tau^*}$	$p_{F1^{\tau^*}}$	$p_{F1^{\tau^*}}^{\mathrm{Holm}}$
ARF	+0.1375	1.9×10^{-6}	1.1×10^{-5}	+0.1177	1.3×10^{-5}	5.3×10^{-5}
Streaming HGB	+0.1341	2.7×10^{-5}	5.3×10^{-5}	+0.0772	1.9×10^{-5}	5.7×10^{-5}
iGBDT	+0.0672	3.6×10^{-5}	5.3×10^{-5}	+0.0488	1.0×10^{-4}	1.0×10^{-4}
DWM	+0.1108	3.8×10^{-6}	1.1×10^{-5}	+0.0908	3.6×10^{-5}	7.2×10^{-5}
Learn++.NSE	+0.1167	1.9×10^{-6}	1.1×10^{-5}	+0.0931	1.9×10^{-6}	1.1×10^{-5}
AUE2	+0.1289	1.9×10^{-6}	1.1×10^{-5}	+0.0911	1.9×10^{-6}	1.1×10^{-5}

C.2 Q2–Q4: Component Ablation Tests

The ablation family contains six planned paired contrasts. Hard voting and removal of the temporal-expert component use directional degradation alternatives; raw precision and removal of the base expert use two-sided alternatives; the Wilson–EWMA accuracy and recent-N variants retain their planned directional contrasts. Holm correction is applied over all six comparisons.

Hard voting keeps expert selection and w_{base} unchanged but thresholds each member probability at 0.5,

$$v_i = \mathbf{1}[p_i \geq 0.5], \qquad s_i^{\mathrm{hard}} = \frac{w_{\mathrm{base}} v_i^{\mathrm{base}} + \sum_{j \in \mathcal{E}_t} v_i^{(j)}}{w_{\mathrm{base}} + |\mathcal{E}_t|}.$$

The `w/o temporal experts` variant disables post-warm-up temporal experts together with their later reassessment, selection, and contribution to fusion, leaving the frozen base expert for prediction. Raw precision uses cumulative $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ over available competence feedback without the Wilson correction or recency weighting. Wilson–EWMA (accuracy) retains $\gamma = 0.85$ and $z = 1.96$ but defines successes as $\mathrm{TP} + \mathrm{TN}$ and failures as $\mathrm{FP} + \mathrm{FN}$. Uniform recent-$N$ selects

up to the N eligible experts with the largest training-week indices. The `w/o base expert` variant sets $w_{\text{base}} = 0$ and renormalizes soft fusion over selected temporal experts; if none are eligible, the prediction score defaults to 0.5.

Table 9: Paired Wilcoxon analysis for the Q2–Q4 MEDT component ablations ($n = 19$, Holm family $m = 6$). ΔPR denotes Full MEDT minus the corresponding variant.

Contrast	ΔPR	Alternative	p_{raw}	p_{Holm}
Hard voting	+0.1023	Full $>$ variant	1.9×10^{-6}	1.1×10^{-5}
w/o temporal experts	+0.0436	Full $>$ variant	0.0006	**0.0029**
Raw precision	-0.0154	two-sided	0.1167	0.4667
Wilson–EWMA (accuracy)	-0.0035	Full $>$ variant	0.3718	1.0000
Uniform recent-N	-0.0005	Full $>$ variant	0.6471	1.0000
w/o base expert	-0.0099	two-sided	0.7680	1.0000

Only hard voting and removal of the temporal-expert component yield Holm-corrected $p < 0.05$ under the planned within-stream paired tests. The remaining contrasts do not establish an independent advantage for the default competence, selection, or anchor choices.

C.3 Q3: MEDT Parameter Sensitivity

Nine non-default settings vary label delay L, committee size N, or EWMA decay γ one factor at a time. All use the same 19 prediction weeks and are compared with the default using paired two-sided Wilcoxon tests with Holm correction.

Table 10: MEDT parameter sensitivity over 19 matched prediction weeks. ΔPR denotes Full MEDT minus the corresponding setting. All non-default contrasts are paired two-sided Wilcoxon tests with Holm correction over the nine planned comparisons.

Group	Setting	PR-AUC	ΔPR	p_{raw}	p_{Holm}
Label delay L	L=0	0.4589	+0.0039	0.2413	1.0000
	L=1	0.4649	-0.0021	0.5949	1.0000
	L=2 (default)	0.4628	—	—	—
	L=3	0.4620	+0.0008	0.1956	1.0000
Committee size N	N=2	0.4605	+0.0022	0.2935	1.0000
	N=3	0.4597	+0.0031	0.2579	1.0000
	N=5 (default)	0.4628	—	—	—
EWMA decay γ	γ=0.80	0.4628	+0.0000	1.0000	1.0000
	γ=0.85 (default)	0.4628	—	—	—
	γ=0.90	0.4631	-0.0003	1.0000	1.0000
	γ=0.95	0.4631	-0.0003	1.0000	1.0000
	γ=1.00	0.4632	-0.0005	0.8750	1.0000

No significant differences are detected among the tested settings after Holm correction. This does not establish statistical equivalence and should not be interpreted as evidence that label delay is irrelevant.

D Controlled Temporal Diagnostics and W26 Protocol Checks

These analyses extend the main evidence with secondary temporal diagnostics. They are separated from the primary statistical support above: isolated single-perturbation runs describe temporal behavior under controlled transformations, while mixed-schedule traces examine committee membership. They do not establish recurrence detection, individual expert causality, or strategic attacker behavior.

D.1 W26 Phase-Wise Performance

For isolated runs, the five reporting phases are Pre-change (weeks 7–9), First change (10–13), Post-change 1 (14–17), Second perturbation (18–21), and Post-change 2 (22–25). The post-change phases follow removal of the preceding perturbation and do not imply recovery to pre-change performance.

Table 11: Phase-wise PR-AUC averaged across the four isolated W26 perturbation runs. Phase windows are weeks 7–9, 10–13, 14–17, 18–21, and 22–25, respectively.

Method	Pre-change	First change	Post-change 1	Second perturbation	Post-change 2
ARF	0.3587	0.3482	0.3410	0.2992	0.2913
Streaming HGB	0.5093	0.4397	0.3815	0.2493	0.1626
iGBDT	0.4239	0.4429	0.4442	0.4146	0.4039
AUE2	0.3310	0.3677	0.3657	0.3241	0.3432
MEDT	0.5135	0.4954	0.4666	0.4631	0.4098

Across the four isolated perturbation runs, MEDT has the highest PR-AUC among the reported methods in all five phase averages, while its performance still declines across repeated controlled changes. The result is therefore descriptive of retained post-change performance rather than complete recovery.

D.2 Perturbation-Specific Final Post-Change PR-AUC

Table 12: Pre-change and final post-change PR-AUC under the four isolated W26 perturbations. The final row reports MEDT's final/pre-change ratio; this is a descriptive retained-performance measure, not a recovery rate.

Method / summary	Pre-change	Amount Mimicry	Credential Churn	Modus Operandi Shift	Traffic Reordering
ARF	0.3587	0.3113	0.2669	0.2790	0.3079
Streaming HGB	0.5093	0.1251	0.1143	0.2457	0.1651
iGBDT	0.4239	0.4114	0.4118	0.3985	0.3940
AUE2	0.3310	0.3412	0.3474	0.3382	0.3462
MEDT	0.5135	0.4073	0.4162	0.4018	0.4139
MEDT final/pre ratio	—	79.3%	81.0%	78.2%	80.6%

Absolute final post-change performance and MEDT's final/pre-change ratio describe different properties. iGBDT slightly exceeds MEDT after Amount Mimicry and Credential Churn, whereas MEDT is higher after Modus Operandi Shift and Traffic Reordering. MEDT also exceeds ARF, Streaming HGB, and AUE2 after all four perturbations. The ratio is descriptive and is not the causal effect of temporal-expert retention.

D.3 Q2: Temporal Expert Selection Traces

Stored traces provide representative mechanism-level examples rather than an exhaustive selection-frequency analysis. Table 13 lists the two examples used here.

Table 13: Representative examples from the stored temporal-expert selection traces; the table does not summarize selection frequency across all weeks.

Context	Later evaluation point	Earlier temporal experts selected
Mixed main evaluation	Week 25	Week 10 remains in Top-5
Amount Mimicry isolated run	Weeks 18–21	Weeks 10, 9, and 7 remain selected

These observations establish continued selection of earlier temporal experts, not their individual causal contribution to PR-AUC or the recurrence of a specific fraud mode.

D.4 W26 Protocol Sensitivity Checks

These diagnostics vary the controlled evaluation schedule rather than MEDT's internal parameters. The span$=4$ setting changes perturbation duration while keeping $L = 2$; the fast-change setting jointly uses span$=1$ and $L = 1$ and therefore does not isolate either factor. The comparisons are exploratory and outside the planned MEDT parameter-sensitivity family.

Table 14: MEDT PR-AUC under selected alternative W26 protocol settings. Two-sided p-values are exploratory paired comparisons with the default and are outside the MEDT parameter-sensitivity Holm family.

Protocol setting	PR-AUC	ΔPR (Default$-$alternative)	Two-sided p
span$=2$, $L = 2$ (default)	0.4628	0.0000	—
span$=4$, $L = 2$	0.4912	-0.0284	0.9032
Fast-change: span$=1$, $L = 1$	0.4799	-0.0171	0.6794

No difference is detected for either selected alternative in these exploratory comparisons. This should not be interpreted as evidence of protocol invariance.

D.5 Credential Churn: Representative Weekly PR-AUC Trajectory

Figure 2 visualizes Credential Churn as a representative isolated trajectory; quantitative conclusions above use all four perturbation runs.

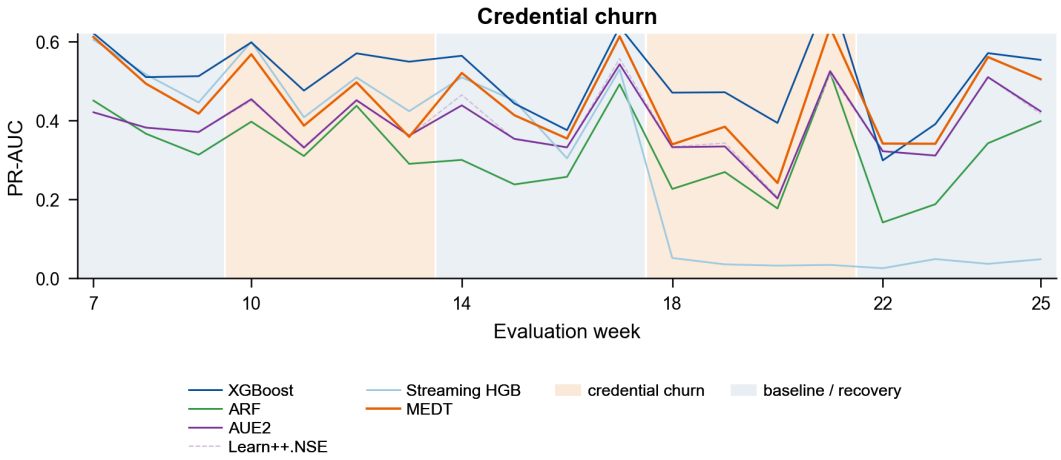


Figure 2: Weekly PR-AUC under the isolated Credential Churn perturbation across the controlled change and delayed-supervision phases.

E Computational and Reproducibility Notes

This section records computational measurements and the execution environment. These details support reproducibility but are not part of MEDT's methodological contribution.

E.1 Computational Characteristics

E.2 Environment and Artifact

The supplementary artifact provides W26 construction code, MEDT and baseline configurations, evaluation and statistical-test scripts, and reproduction entry points for the reported tables and figures. All experiments use random seed 42. The implementation was verified on an Apple M1 Pro

Table 15: Computational characteristics of representative methods over 19 evaluation weeks, including the XGBoost batch retraining reference. Wall, Train, and Predict are average weekly runtimes; Peak RSS is maximum memory usage.

Method	Wall (s)	Train (s)	Predict (s)	Peak RSS (MiB)
XGBoost (retraining ref.)	11.77	11.71	0.06	6485.3
ARF	41.00	37.92	3.09	7442.6
Streaming HGB	15.02	14.71	0.30	4920.6
iGBDT	8.71	8.56	0.15	6999.2
AUE2	266.31	259.32	4.84	5862.8
MEDT	68.27	56.83	7.96	5333.1

system with 16 GB unified memory; small numerical differences may occur across software and hardware environments because of implementation-dependent numerical behavior.