STRICT ONLINE LEARNING FOR WILD STREAMING DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world streaming data commonly exhibit characteristics such as concept drift, missing values, and heterogeneous features, and such data are referred to as *wild streaming data*. Existing approaches often rely on buffers that retain past instances to support model adaptation, while the implications of data retention, including privacy and data governance concerns, have received limited attention. We investigate a new problem, *Strict Online Learning for Wild Streaming Data*, in which each instance is used once for prediction and update and then discarded. This no-retention requirement makes learning challenging, as the model adapts to evolving patterns one instance at a time without destabilizing previously acquired knowledge. To address this challenge, we propose **OnlineNet**, a fast–slow online neural framework for regulating adaptation under uncertain single-instance evidence. Its slow pathway contextualizes the current observation using accumulated stream context, while its fast pathway captures instantaneous evidence. Bounded Adaptive Modulation coordinates their interaction by controlling how instantaneous evidence influences the slow representation. To stabilize single-instance learning, we introduce Online Prior Adjustment based on a prequential class prior estimated solely from previously revealed labels. Extensive experiments across 81 streaming conditions derived from 27 public datasets demonstrate that OnlineNet achieves the best aggregate predictive performance across all four cumulative metrics. It also maintains high computational efficiency while using only 0.5% of the parameters required by a competitive baseline. The source code and datasets are submitted as Supplementary Materials for Reproducibility.

1 INTRODUCTION

Streaming data Nguyen et al. (2015) arise continuously in many real-world applications, where instances arrive sequentially and often need to be processed as soon as they become available. Online learning Hoi et al. (2021) provides a natural paradigm for such scenarios by incrementally updating the model with each new instance, allowing it to make predictions and adapt as the stream evolves. Unlike conventional batch learning, which assumes access to a complete dataset before training, online learning updates the model directly as new instances arrive, making it particularly well suited to real-time and evolving environments Ortiz Díaz et al. (2025).

To better reflect real-world deployment conditions, online learning has increasingly been studied in non-stationary environments Fahy et al. (2022); He et al. (2023). Real-world streaming data commonly exhibit characteristics such as concept drift, missing values, and heterogeneous features, and such data are referred to as *wild streaming data*. Existing approaches Zhuo et al. (2025) often rely on buffers that retain past instances to support model adaptation, while the implications of data retention have received limited attention Valkanas et al. (2025). In privacy-sensitive applications, streaming data may contain sensitive behavioral, contextual, or physiological information. Persistently retaining such data may raise privacy and data governance concerns Yang et al. (2022), as elaborated in Appendix A. We therefore consider a stricter protocol beyond the standard online learning setting Hoi et al. (2021), in which each instance is used for the current prediction and model update and is then discarded without subsequent replay or reuse.

With historical data unavailable for future reuse, the learner cannot rely on revisiting past instances to maintain previously acquired knowledge, placing greater importance on preserving that knowl-

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.

edge within the model. Deep neural networks Sahoo et al. (2018) remain attractive in this setting because of their strong representation capacity. Nevertheless, adapting deep predictive models from individual instances remains challenging Aljundi et al. (2019). Conventional deep learning typically benefits from mini-batch training and repeated optimization over available data Sekhari et al. (2021); Lin & Rosasco (2017), whereas strict online learning must update the model using the current instance alone. Wild streaming data further complicate this learning process, as an incoming instance may be incomplete while the underlying predictive relationship continues to evolve (Wang et al., 2026b). Consequently, the learner must update its model from limited evidence whose long-term relevance may be difficult to determine upon arrival.

We formulate this problem as *Strict Online Learning for Wild Streaming Data*, with its central challenge arising from the *stability-plasticity dilemma* Grossberg (1980) under uncertain single-instance evidence. A central question is how strongly each arriving observation should influence the model. Rapid adaptation is necessary when the stream reveals meaningful changes in the predictive structure, but aggressive updates driven by transient or incomplete evidence may compromise useful knowledge. However, overly conservative updates may prevent the learner from tracking such changes. The tension becomes consequential when observations cannot be revisited, since inappropriate updates cannot be corrected by replaying historical observations. Class frequencies may also evolve over the stream, causing more frequently observed classes to exert greater cumulative influence through sequential single-instance updates. The goal is therefore to regulate adaptation at the instance level while accounting for class-frequency effects, allowing the model to respond rapidly to meaningful changes while preserving and selectively updating useful knowledge.

To achieve this objective, we draw inspiration from the Complementary Learning Systems (CLS) theory McClelland et al. (1995), which highlights the value of complementary processes for rapid acquisition and stable knowledge consolidation. Accordingly, we propose **OnlineNet**, a fast–slow online neural framework tailored to wild streaming data. The slow pathway contextualizes the current observation against accumulated stream context using causally maintained statistics, while the fast pathway directly captures the current feature values and their availability pattern. To coordinate these sources, we introduce Bounded Adaptive Modulation (BAM), which imposes an asymmetric inductive bias by allowing instantaneous evidence from the fast pathway to modulate the slow representation within a bounded range, enabling adaptation while limiting disruption. We further introduce Online Prior Adjustment (OPA), which estimates the class prior from previously revealed labels to account for class-frequency effects during sequential single-instance updates.

In summary, our work makes the following contributions. First, we investigate strict online learning for wild streaming data, a practically relevant yet underexplored setting in existing streaming and online learning research. Second, motivated by the CLS theory, we propose OnlineNet, a fast–slow online neural framework that separates accumulated stream context from instantaneous evidence and coordinates them through BAM. We further introduce OPA to account for class-frequency effects in sequential single-instance updates using only previously revealed labels. Third, we conduct extensive experiments across 81 streaming conditions derived from 27 public datasets, demonstrating strong predictive performance and computational efficiency under diverse feature-availability and stream conditions.

2 RELATED WORK

Additional related work regarding privacy-oriented learning paradigms is provided in Appendix B.

2.1 ONLINE LEARNING FOR WILD STREAMING DATA

Early online learning algorithms such as OGD Zinkevich (2003) and FTRL McMahan et al. (2013) provide theoretical guarantees under standard online optimization settings. However, these guarantees are often established under simplified assumptions that may not fully capture the complexity of real-world streams. In practice, streaming data may exhibit characteristics such as concept drift caused by changes in user preferences or environmental conditions Gaudreault & Branco (2024), missing features resulting from sensor failures Huang & Evans (2025), and heterogeneous features consisting of both numerical and categorical attributes He et al. (2021). Streams exhibiting such characteristics are commonly referred to as *wild streaming data* Wang et al. (2026b). Recent studies

have introduced dedicated adaptation mechanisms for such wild streaming scenarios. For example, OVFM He et al. (2021) uses a buffer of recent instances to estimate Gaussian copula mappings and feature dependencies for missing-feature reconstruction. OL-MDISF Zhuo et al. (2025) further handles distribution drift and incomplete supervision by maintaining an adaptive sliding window to detect distributional changes. Subsequently, OLWDF Wang et al. (2026b) achieves latent space estimation and geometric adaptation under delayed supervision by maintaining a multi-granularity buffer of recent observations.

Despite their effectiveness, these approaches implicitly assume that recent historical observations can be accessed repeatedly. In contrast, this paper extends research on wild streaming data to a strict online protocol, where each instance is used once for prediction and update and then discarded.

2.2 Online Deep Learning

Online deep learning Sulaiman et al. (2025); Ortiz Díaz et al. (2025) extends deep representation learning from static batch settings to streaming environments, enabling models to update incrementally as new instances arrive without repeated batch retraining. A seminal contribution in this domain is the ODL framework Sahoo et al. (2018), which attaches classifiers to different hidden layers and combines their predictions through Hedge Backpropagation Freund & Schapire (1995), allowing the model to exploit representations at different depths during streaming learning. Subsequent studies further extended the architectural scope of ODL. Auxiliary Network Agarwal et al. (2022) introduces feature-specific auxiliary components by distinguishing consistently available base features from auxiliary features, while Aux-Drop Agarwal et al. (2023) reduces dependence on auxiliary features with uncertain availability. OLD^{3}S Lian et al. (2024) further introduces a shared latent subspace between old and newly arriving features, enabling adaptation to changing feature spaces. More recently, MODL Valkanas et al. (2025) improves online deep learning efficiency by replacing Hedge Backpropagation with a hybrid optimization strategy that combines gradient-based neural learning with recursive closed-form updates. HI^2 Agarwal et al. (2025a) introduces a model-agnostic image transformation that maps variable-dimensional inputs into fixed-size representations, enabling online learning to leverage pretrained vision models and the prior knowledge encoded in them.

Although these studies expand the applicability of deep models to streaming environments, the stability–plasticity dilemma in wild streaming data with non-stationary characteristics remains insufficiently explored. In this paper, we address this challenge through complementary fast and slow pathways that coordinate accumulated stream context with instantaneous evidence.

2.3 Stability-Plasticity in Online Adaptation

The *stability-plasticity dilemma* concerns how a learner can absorb new information without overwriting previously acquired knowledge Grossberg (1980). Complementary Learning Systems (CLS) theory provides a conceptual framework for balancing rapid adaptation and gradual knowledge consolidation McClelland et al. (1995). Recent CLS-inspired approaches have explored this principle in continual learning Pham et al. (2024); Zhao et al. (2024) and online time-series forecasting Pham et al. (2022); Lau et al. (2025). However, existing CLS-inspired approaches have not been developed specifically for online adaptation to wild streaming data and commonly rely on replay, auxiliary memory, or accumulated temporal context to support stable adaptation. In this paper, we extend the CLS principle to strict online learning for wild streaming data, where rapid adaptation and knowledge preservation must be achieved without retaining historical observations.

3 Problem Statement

We investigate online learning for wild streaming data under the strict online protocol. At round $t \in \{1, \ldots, T\}$, the learner receives an input $\mathbf{x}_t \in \mathbb{R}^d$ with heterogeneous features, including continuous, ordinal, and Boolean attributes, together with an availability mask $\mathbf{m}_t \in \{0,1\}^d$, where $m_{t,j} = 1$ indicates that feature j is observed. We use the zero-filled input $\widetilde{\mathbf{x}}_t = \mathbf{m}_t \odot \mathbf{x}_t$ while retaining $\mathbf{m}_t$ as an explicit input, so that an unavailable feature is distinguishable from an observed zero. At each round, $(\mathbf{x}_t, \mathbf{m}_t, y_t)$ follows a distribution P_t that may vary over time, with $y_t \in \mathcal{Y}$ and $\mathcal{Y} = \{0, 1\}$ in our experiments.

Learning follows the prequential test-then-train procedure Dawid (1984); Gama et al. (2013). At the beginning of round t, the learner uses the current model F_{Θ_t} to produce a prediction $\hat{y}_t = F_{\Theta_t}(\widetilde{\mathbf{x}}_t, \mathbf{m}_t)$ before y_t is revealed. Once the label is revealed, the model performs a single supervised model update to obtain Θ_{t+1}. Under the strict online protocol, each instance is used once for prediction and update and then discarded, with only model parameters and compact summary statistics carried forward.

4 METHOD

OnlineNet is designed to regulate learning from uncertain single-instance evidence under the strict online protocol. Its core idea is to separate accumulated stream context from instantaneous evidence through complementary slow and fast pathways. The slow pathway contextualizes the current observation using causally maintained stream statistics, while the fast pathway directly encodes the current observed values and their availability pattern. Bounded Adaptive Modulation (BAM) then allows instantaneous evidence from the fast pathway to modulate the slow representation within a bounded, non-attenuating range. Finally, Online Prior Adjustment (OPA) incorporates a prequential class prior estimated from previously revealed labels into the online training objective. The overall procedure is summarized in Appendix C.2.

4.1 COMPLEMENTARY FAST–SLOW REPRESENTATION LEARNING

Causal Stream Contextualization Under strict online learning, past instances are discarded after use and therefore cannot be revisited to contextualize future inputs. To retain such context without storing raw instances, we maintain a compact causal stream state $\mathcal{C}_{t-1}$ through recursively updated feature-wise statistics. The current observation is then contextualized with respect to this state:

$$\mathbf{z}_t = \mathbf{m}_t \odot \mathrm{Norm}\left(\widetilde{\mathbf{x}}_t; \mathcal{C}_{t-1}\right). \tag{1}$$

Here, $\mathbf{z}_t$ denotes the causally normalized input to the slow pathway, and $\mathrm{Norm}(\cdot)$ performs feature-wise normalization using $\mathcal{C}_{t-1}$. This provides the slow pathway with a stable history-conditioned reference, while the mask prevents unavailable features from introducing artificial evidence. After round t, the observed feature values are incorporated into the stream state to obtain $\mathcal{C}_t$, which is used from the next round onward. As this cumulative state evolves gradually with the stream, it remains relatively stable under short-term fluctuations, while changes in the current observation are reflected more directly by the fast pathway. The full mask-aware causal normalization procedure is detailed in Appendix C.1.

Instantaneous Evidence Encoding While the slow pathway contextualizes each observation using the causal stream state, this process may attenuate information specific to the current observation. We therefore introduce a fast pathway that operates directly on the current observed values and their availability pattern, preserving complementary instantaneous evidence:

$$\mathbf{f}_{0,t} = [\widetilde{\mathbf{x}}_t; \mathbf{m}_t], \qquad \mathbf{f}_{\ell,t} = \mathrm{ReLU}\left(\mathbf{W}_\ell^f \mathbf{f}_{\ell-1,t} + \mathbf{b}_\ell^f\right), \quad \ell = 1, \ldots, L. \tag{2}$$

Here, $\mathbf{f}_{0,t}$ is the input to the fast pathway, and $\mathbf{f}_{\ell,t}$ is its representation at layer ℓ.

4.2 BOUNDED ADAPTIVE MODULATION

Directional Fast-to-Slow Interaction The slow and fast pathways provide complementary signals but play asymmetric roles. The slow pathway contextualizes the current observation against accumulated stream context, whereas the fast pathway captures instantaneous evidence directly from the current input. Symmetric fusion, such as addition or concatenation, imposes no explicit directional prior and therefore obscures the distinct roles of the slow and fast pathways. We therefore impose an asymmetric inductive bias in which the fast pathway modulates the slow representation, allowing it to adapt to instantaneous evidence while retaining accumulated stream context. Accordingly, at layer ℓ, the slow pathway first computes the unmodulated slow representation:

$$\mathbf{s}_{\ell,t} = \mathrm{ReLU}\left(\mathbf{W}_\ell^s \mathbf{h}_{\ell-1,t} + \mathbf{b}_\ell^s\right), \qquad \mathbf{h}_{0,t} = \mathbf{z}_t. \tag{3}$$

Bounded Non-Attenuating Modulation A key design consideration is how instantaneous evidence should influence the slow representation. A larger response from the fast pathway induces greater modulation, whereas a limited response provides no reliable basis for attenuation and may otherwise conflate insufficient evidence with negative evidence. We therefore adopt a non-attenuating modulation principle that allows the fast pathway to amplify the slow representation while defaulting to preservation under limited responses. Using the multiplicative identity as the neutral state, we restrict the modulation to non-negative deviations from it. Accordingly, the modulation gate at layer ℓ is defined as:

$$\mathbf{g}_{\ell,t} = \mathbf{1} + \lambda \left(2\sigma(\mathbf{f}_{\ell,t}) - \mathbf{1}\right), \qquad \lambda > 0, \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The gate then modulates the slow representation as:

$$\mathbf{h}_{\ell,t} = \mathbf{s}_{\ell,t} \odot \mathbf{g}_{\ell,t}. \tag{5}$$

Here, $\mathbf{h}_{\ell,t}$ denotes the modulated representation passed to the next paired layer or, at the final layer, to the classifier. Since the fast pathway uses ReLU activations, $\mathbf{f}_{\ell,t} \geq \mathbf{0}$, yielding $2\sigma(\mathbf{f}_{\ell,t}) - \mathbf{1} \in [0, 1)$ and thus $\mathbf{1} \leq \mathbf{g}_{\ell,t} < (1+\lambda)\mathbf{1}$. The gate is therefore bounded and non-attenuating, with λ controlling the maximum amplification. While BAM restricts $\lambda > 0$, we also consider negative values as counterfactual variants that permit attenuation in Section 5.4. The modulation is applied at each paired layer, and the resulting representation $\mathbf{h}_{L,t}$ is passed to a linear classifier to produce the raw prediction logits:

$$\mathbf{o}_t = \mathbf{W}^{\text{out}}\mathbf{h}_{L,t} + \mathbf{b}^{\text{out}}. \tag{6}$$

4.3 Prior-Adjusted Online Training

While the preceding representation mechanism regulates uncertainty associated with the current observation, strict online learning introduces a separate challenge because the class prior is not known in advance and must be estimated progressively as the stream unfolds. The label for each arriving instance is revealed only after prediction and then drives a single supervised model update, so the learner cannot use a global class prior computed from the full stream. Inspired by Menon et al. (2021), we incorporate a prequential class prior estimated solely from previously revealed labels into the strict online training objective. We refer to this mechanism as Online Prior Adjustment (OPA) and let $N_{t-1,c}$ denote the number of labels from class c revealed before round t. With Laplace smoothing, the prequential prior is defined as:

$$\widehat{\pi}_{t,c} := \frac{N_{t-1,c} + \alpha}{\sum_{k\in\mathcal{Y}} N_{t-1,k} + \alpha|\mathcal{Y}|}, \qquad c \in \mathcal{Y}, \quad \alpha = 1. \tag{7}$$

Laplace smoothing ensures a well-defined prior even when a class has not yet appeared in the observed stream. To preserve the prequential ordering, the prediction is obtained from the raw logits $\mathbf{o}_t$, while the prior is used only after y_t is revealed to construct the prior-adjusted training logits:

$$\overline{o}_{t,c} = o_{t,c} + \tau \log \widehat{\pi}_{t,c}, \qquad c \in \mathcal{Y}, \tag{8}$$

where $\overline{o}_{t,c}$ is the prior-adjusted training logit for class c, and $\tau \geq 0$ is the adjustment coefficient. The per-round loss is then defined as:

$$\mathcal{L}_t := -\log \frac{\exp\left(\overline{o}_{t,y_t}\right)}{\sum_{c\in\mathcal{Y}} \exp\left(\overline{o}_{t,c}\right)}. \tag{9}$$

For a class with a smaller prequential prior, the adjustment introduces a larger negative offset to its training logit, requiring the model to learn a larger raw-logit margin and thereby helping counteract the frequency effects accumulated through sequential single-instance updates. Only after the supervised update is y_t incorporated into the class counts, preserving the prequential ordering by allowing the current label to affect only subsequent priors.

5 EXPERIMENTS

Our experiments aim to test the following hypotheses. **(H1)** OnlineNet achieves competitive predictive performance against representative classical and deep online learning methods under different

feature availability conditions. **(H2)** OnlineNet maintains competitive online performance as instances arrive sequentially and adapts as the stream evolves. **(H3)** OnlineNet provides favorable computational efficiency in terms of runtime and trainable parameter count. **(H4)** The main components of OnlineNet provide complementary contributions to its overall performance. **(H5)** OnlineNet remains robust to the modulation and adjustment coefficients used by BAM and OPA. Due to space constraints, the main text presents the key settings and results, with additional details in the Appendix and exploratory analyses in Appendix F.

5.1 Experimental Settings

Datasets We evaluate OnlineNet on 27 datasets, comprising 15 real-world datasets from the UCI repository Asuncion et al. (2007) and 12 synthetic datasets generated with the MOA Bifet et al. (2023). The real-world datasets span 11 application domains. The synthetic datasets explicitly simulate abrupt and gradual concept drift by providing clearly defined drift positions, thereby enabling precise evaluation. Across all datasets, the number of instances ranges from 569 to 50,000, the number of features ranges from 3 to 499, and the majority-class percentage ranges from 50.32% to 76.19%. These datasets therefore provide diverse evaluation settings in terms of stream size, feature dimensionality, and class distribution. Their detailed descriptions are provided in Appendix D.1.

Following prior work He et al. (2021); Valkanas et al. (2025); Agarwal et al. (2023; 2025a), we generate availability masks using a feature observation probability p. Each feature is initially sampled as observed with probability p. If this produces an all-zero mask, one feature is retained at random. We consider $p \in \{0.75, 0.50, 0.25\}$, corresponding to mild, moderate, and severe feature unavailability, respectively. Combining the 27 datasets with the three values of p yields 81 streaming conditions.

Metrics We report Accuracy, AUROC, AUPRC, Balanced Accuracy, and Sliding-Window Error. Detailed definitions are deferred to Appendix D.2.

Baselines We compare the proposed method against eight baselines, categorized into two paradigms: (1) *Classical Online Learning*: ORF^{3}V Schreckenberger et al. (2023), OLIFL You et al. (2024), OLVF Beyazit et al. (2019), and OLWDF Wang et al. (2026b). (2) *Online Deep Learning*: HBP Sahoo et al. (2018), Aux-Drop Agarwal et al. (2023), MODL Valkanas et al. (2025), and HI^2 Agarwal et al. (2025a). Detailed descriptions of the baselines are provided in Appendix D.3.

Implementation Details We report a single run for OLVF and OLIFL because their outputs are deterministic and seed-invariant. For all other methods, we report the mean $\pm$ standard deviation over five runs on real-world datasets using seeds $\{0, 1, 2, 3, 4\}$ and three runs on synthetic datasets using seeds $\{0, 1, 2\}$. We provide further details in Appendix D.4.

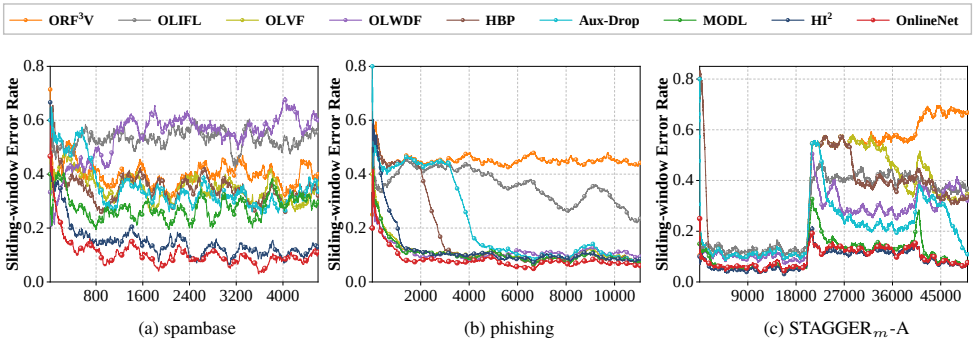
5.2 Main Results

5.2.1 Cumulative Performance (H1)

Table 1 summarizes the average cumulative performance over 27 datasets under three feature observation probabilities. OnlineNet achieves the best aggregate performance across all evaluation metrics. Among the baselines, HI^2 is consistently the strongest competitor, followed by OLWDF and MODL. As p decreases, predictive performance declines across all methods. Nevertheless, OnlineNet exhibits an increasing relative advantage over HI^2, with the margins in AUROC and AUPRC increasing from 2.17 and 2.24 percentage points at $p = 0.75$ to 3.87 and 4.20 percentage points at $p = 0.25$, respectively. Together, the rank, win/loss, and significance results indicate that OnlineNet's advantage extends across experimental settings and all four cumulative metrics, rather than being limited to a small subset of favorable cases or a single metric. Across the three feature availability ratios, the average reported standard deviations of OnlineNet remain between 0.26% and 0.40% for all metrics, indicating low run-to-run variability. Overall, these results show that OnlineNet retains strong predictive performance as the observed feature space becomes sparse. Detailed experimental and statistical results are available in Appendix E.1.

Table 1: Average results (%) across 27 datasets at three feature observation probabilities p. Each method occupies two rows reporting mean performance and the average reported per-dataset standard deviation across repeated runs. Bold indicate the best results. The symbol "–" denotes unavailable standard deviations for the single-run methods OLVF and OLIFL. Avg. Rank ↓ and Win/Loss are based on Balanced Accuracy over the 81 streaming conditions, with ties omitted. OnlineNet significantly outperforms every baseline for all four cumulative metrics with $p_{\text{adj}} < 0.05$.

Method	$p = 0.25$				$p = 0.50$				$p = 0.75$				Statistics	
	Acc.	AUROC	AUPRC	Bal. Acc.	Acc.	AUROC	AUPRC	Bal. Acc.	Acc.	AUROC	AUPRC	Bal. Acc.	Avg. Rank	Win/Loss
Classical Online Learning														
ORF^{3}V	52.12 (±0.32)	47.08 (±0.23)	47.68 (±0.15)	45.88 (±0.30)	52.55 (±0.31)	46.35 (±0.33)	47.26 (±0.21)	46.03 (±0.31)	52.68 (±0.25)	45.14 (±0.26)	46.51 (±0.18)	46.05 (±0.27)	8.78	0/81
OLIFL	57.52 –	45.24 –	57.99 –	54.62 –	59.75 –	44.24 –	57.02 –	55.73 –	61.30 –	42.58 –	55.94 –	57.42 –	6.09	1/80
OLVF	60.12 –	55.58 –	55.22 –	53.93 –	62.78 –	55.36 –	55.45 –	56.54 –	64.84 –	57.07 –	57.22 –	58.83 –	5.99	0/81
OLWDF	68.95 (±0.07)	71.09 (±0.08)	69.95 (±0.09)	63.67 (±0.08)	72.16 (±0.07)	76.13 (±0.08)	74.65 (±0.09)	67.78 (±0.07)	71.82 (±1.45)	74.58 (±1.71)	74.40 (±1.89)	68.39 (±1.20)	3.59	2/79
Online Deep Learning														
HBP	61.29 (±1.23)	55.75 (±3.13)	55.87 (±2.98)	53.44 (±1.07)	63.31 (±1.27)	59.82 (±3.10)	59.37 (±2.66)	56.27 (±1.10)	65.09 (±1.39)	63.11 (±2.95)	62.55 (±2.54)	58.49 (±1.28)	6.65	0/81
Aux-Drop	62.42 (±1.29)	58.30 (±0.97)	58.05 (±0.81)	55.03 (±0.59)	64.78 (±1.23)	62.66 (±1.02)	61.98 (±1.05)	57.99 (±0.87)	67.74 (±1.29)	66.76 (±1.47)	65.95 (±1.48)	61.71 (±1.07)	6.03	0/81
MODL	67.10 (±0.39)	65.98 (±0.51)	65.76 (±0.75)	60.63 (±0.52)	69.69 (±0.40)	69.54 (±0.64)	69.31 (±0.87)	63.67 (±0.52)	73.56 (±0.55)	74.34 (±0.88)	73.75 (±1.21)	68.22 (±0.81)	4.14	0/81
HI^2	70.11 (±0.52)	72.14 (±0.59)	71.91 (±0.71)	64.32 (±0.61)	75.42 (±0.46)	79.62 (±0.47)	79.50 (±0.53)	70.61 (±0.55)	80.56 (±0.45)	85.49 (±0.36)	85.47 (±0.41)	76.71 (±0.55)	2.67	2/79
OnlineNet	**72.46** (±0.40)	**76.01** (±0.29)	**76.11** (±0.29)	**68.02** (±0.39)	**77.57** (±0.33)	**82.71** (±0.26)	**82.71** (±0.28)	**74.31** (±0.39)	**82.56** (±0.34)	**87.66** (±0.26)	**87.71** (±0.27)	**80.04** (±0.36)	**1.06**	643/5

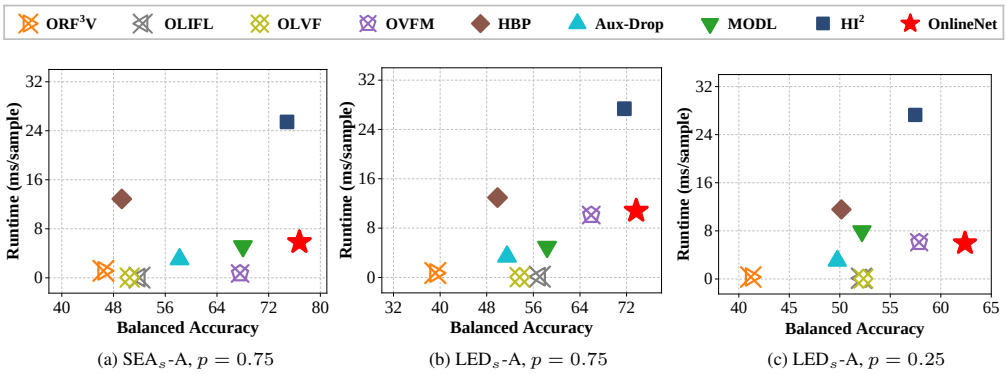


(a) spambase (b) phishing (c) STAGGER$_m$-A

Figure 1: Performance evolution of different methods as instances arrive sequentially at $p = 0.75$.

5.2.2 Online Performance Evolution (H2)

Cumulative metrics summarize overall predictive performance but do not reveal how performance evolves as instances arrive sequentially. We therefore examine the sliding-window error under the strict online protocol. As shown in Figure 1, OnlineNet generally maintains a low-error trajectory throughout the stream. For the real-world datasets in Figure 1(a), the error trajectories exhibit dataset-specific fluctuations. Since the locations of distributional changes are not annotated, individual error peaks should not be directly interpreted as evidence of concept drift. Nevertheless, OnlineNet maintains comparatively low error levels even on streams with stronger fluctuations. The results on the controlled synthetic data further illustrate model behavior under known distributional changes. In particular, on STAGGER$_m$-A, OnlineNet exhibits a comparatively limited error increase around concept transitions and maintains a low-error regime afterward. Overall, these results show that OnlineNet sustains competitive local predictions during continuous updates while adapting to newly arriving information, showing that OnlineNet maintains comparatively low local error while adapting to changes in the stream. More experimental results are provided in Appendix E.2.



(a) SEA$_s$-A, $p = 0.75$ (b) LED$_s$-A, $p = 0.75$ (c) LED$_s$-A, $p = 0.25$

Figure 2: Performance–efficiency trade-off in terms of balanced accuracy and per-instance runtime.

Table 2: Parameter comparison on spam. Relative count is the ratio to OnlineNet's parameter count.

	HBP	Aux-Drop	MODL	HI^2	**OnlineNet**
Parameters ↓	≈ 163,000	≈ 1,660,000	≈ 1,357,000	≈ 21,285,000	≈ **104,000**
Relative size	≈ 1.57×	≈ 15.96×	≈ 13.05×	≈ 204.66×	≈ **1.00**×

5.2.3 Computational Efficiency (H3)

While predictive performance remains essential, practical deployment of online models also requires efficient computation for processing streaming data. Figure 2 compares balanced accuracy and per-instance runtime on three representative synthetic streams. OnlineNet achieves the highest balanced accuracy while requiring substantially less runtime than HI^2, the strongest competing baseline in predictive performance. Several classical online methods require less runtime but achieve lower predictive performance. Moreover, Figure 2(a)–(b) shows that OLWDF has substantially higher per-instance runtime on LED$_s$-A than on SEA$_s$-A, whereas OnlineNet's runtime changes little between the two streams. This comparison suggests that OnlineNet is less sensitive to increased input dimensionality, making it potentially better suited to higher-dimensional streaming data.

Beyond runtime efficiency, we compare the number of trainable parameters. Table 2 reports the trainable parameter count on spam, the highest-dimensional dataset in our benchmark. With only 104K trainable parameters, OnlineNet achieves competitive predictive performance while using substantially fewer trainable parameters than all baselines, indicating high parameter efficiency. The time and space complexity analyses are provided in Appendix E.3.

5.3 Ablation Studies (H4)

We assess the contribution of each component in OnlineNet using six variants. (1) **w/o Causal** removes causal normalization based on causally maintained stream statistics. (2) **w/o Fast**$_{v1}$ removes the fast pathway and fixes $\mathbf{g}_{\ell,t} = \mathbf{1}$. (3) **w/o Fast**$_{v2}$ generates $\mathbf{g}_{\ell,t}$ from $\mathbf{s}_{\ell,t}$ using a lightweight gating network without an independent fast pathway. (4) **w/o BAM**$_{v1}$ replaces BAM with concatenation and linear projection. (5) **w/o BAM**$_{v2}$ replaces BAM with element-wise addition. (6) **w/o OPA** removes the prior adjustment and uses standard cross-entropy.

We report the ablation results in Table 3 shows that **w/o Causal** causes the largest overall degradation, indicating that causal stream statistics contribute most to aggregate performance. **w/o Fast**$_{v1}$ and **w/o Fast**$_{v2}$ perform similarly, showing that generating the gate from the slow representation alone offers little benefit over fixing it to $\mathbf{1}$. Both variants remain below OnlineNet on all four metrics, which is consistent with the independent fast pathway supplying instantaneous evidence. **w/o BAM**$_{v2}$ performs better than **w/o BAM**$_{v1}$, but neither matches OnlineNet, suggesting that direct combination cannot reproduce the benefit of BAM's bounded fast-to-slow modulation. Finally, the results for **w/o OPA** suggest that OPA regulates class-frequency effects rather than improving all predictive metrics uniformly.

Table 3: Ablation results averaged over six datasets at $p = 0.75$. Bold indicates the best result in each row. Detailed per-dataset results are provided in Appendix E.4.

Metric	w/o Causal	w/o Fast$_{v1}$	w/o Fast$_{v2}$	w/o BAM$_{v1}$	w/o BAM$_{v2}$	w/o OPA	OnlineNet
Accuracy ↑	80.642	82.973	82.943	82.462	82.665	**83.363**	83.323
AUROC ↑	80.212	85.040	84.927	81.995	83.283	85.115	**85.212**
AUPRC ↑	86.558	90.472	90.387	88.408	88.968	90.625	**90.643**
Bal. Acc. ↑	73.427	76.772	76.753	75.028	76.243	76.092	**77.422**

5.4 Hyperparameter Analysis (H5)

5.4.1 Modulation Coefficient

We investigate the effect of the modulation coefficient λ, which controls the influence of instantaneous evidence on the slow representation. Negative values are included as counterfactual variants that permit attenuation, whereas BAM itself uses $\lambda > 0$. As shown in Figure 3, negative values lead to lower balanced accuracy and larger variance, supporting the non-attenuating design. For positive λ, performance improves within a moderate range, while further increasing the modulation coefficient brings limited additional gains. These results suggest that preserving the slow representation while allowing bounded amplification provides a more stable interaction between instantaneous evidence and accumulated stream context. More results are provided in Appendix E.5.1.

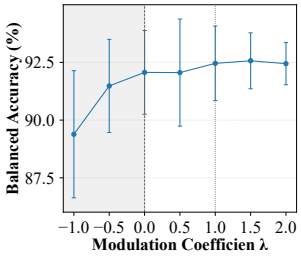


Figure 3: Effect of modulation coefficient λ on wdbc. Error bars indicate standard deviations.

5.4.2 Adjustment Coefficient

We examine the adjustment coefficient τ, which controls the contribution of the prequential class prior to the training objective. Figure 4 shows that increasing τ improves balanced accuracy, while overall accuracy remains stable under moderate adjustment and declines under the strongest setting. This pattern suggests that prior adjustment can mitigate class-frequency effects, although excessive adjustment may reduce overall accuracy. We therefore use $\tau = 0.25$ as a practical compromise rather than treating it as a dataset-independent optimum. Appendix E.5.2 reports results on additional datasets and provides tuning guidance.

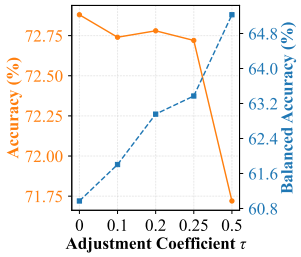


Figure 4: Effect of adjustment coefficient τ on german.

6 CONCLUSION

We investigate strict online learning for wild streaming data and propose OnlineNet, a fast–slow online neural framework designed for learning from uncertain single-instance evidence. Extensive experiments across 81 streaming conditions derived from 27 public datasets demonstrate that OnlineNet achieves the best aggregate predictive performance across all four cumulative metrics. It also maintains high computational efficiency and uses only 0.5% of the parameters of the strongest competing baseline. Despite these advantages, several limitations remain. First, providing regret guarantees for nonlinear online neural networks such as OnlineNet remains challenging, as existing theoretical analyses often rely on assumptions regarding model structures and optimization processes Gao et al. (2019) that may not align with realistic non-stationary streaming scenarios. Establishing theoretical guarantees under sufficiently mild and realistic assumptions therefore remains an important direction for future research. Second, the no-retention requirement does not constitute a formal privacy guarantee, since historical information may still be implicitly encoded in model parameters. Future work will explore integrating machine unlearning Bourtoule et al. (2021) into online learning on streaming data, aiming to remove the influence of previously observed data from continuously updated models.

AI USE STATEMENT

During the preparation of this manuscript, ChatGPT, a large language model, was used to assist with language editing, including grammar correction, sentence refinement, and clarification of wording where necessary. Its use was limited to improving the readability and presentation of the manuscript. All research ideas, methodologies, experimental designs, and analyses presented in this work were developed and conducted by the authors.

ETHICS STATEMENT

Our work investigates strict online learning for wild streaming data under a strict online protocol, where each arriving instance is used once for prediction and model update and then discarded without replay or reuse. This setting reduces the need for persistent retention of historical observations and may therefore be beneficial in privacy- and data-governance-sensitive streaming applications involving behavioral, contextual, or user-related information. However, this protocol should not be interpreted as providing a formal privacy guarantee, since model parameters and internal representations are retained and may still encode information derived from previously observed data. Moreover, deployment in real-world decision-making systems requires careful consideration of potential bias and reliability issues arising from non-stationary distributions, incomplete observations, heterogeneous features, and evolving class frequencies, as these factors may affect different data populations or operating conditions unevenly. OnlineNet is intended as a learning framework for studying prediction and adaptation under strict data-retention constraints rather than as a mechanism for guaranteeing privacy, fairness, or safe decision-making. Responsible deployment should therefore consider data sensitivity, application-specific risks, prediction consequences, and appropriate safeguards.

REPRODUCIBILITY STATEMENT

To facilitate the reproducibility of our work, we provide the source code, dataset preparation scripts, implementation details, and instructions required to reproduce the experiments through an anonymous repository: https://anonymous.4open.science/r/OnlineNet-DFA2. The experimental protocol is described in detail in the main text and Appendix, including the construction of 81 streaming conditions from 27 public datasets using feature availability ratios $p \in \{0.25, 0.50, 0.75\}$, the prequential test-then-train evaluation procedure, evaluation metrics, baseline configurations, and repeated-run protocol. For stochastic methods, we report results averaged over five runs with random seeds $\{0, 1, 2, 3, 4\}$ on real-world datasets and three runs with random seeds $\{0, 1, 2\}$ on synthetic datasets, while deterministic methods are evaluated once.

REFERENCES

Rohit Agarwal, Krishna Agarwal, Alexander Horsch, and Dilip K. Prasad. Auxiliary network: Scalable and agile online learning for dynamic system with inconsistently available inputs. In *International Conference on Neural Information Processing*, pp. 549–561, 2022.

Rohit Agarwal, Deepak Gupta, Alexander Horsch, and Dilip K. Prasad. Aux-drop: Handling haphazard inputs in online learning using auxiliary dropouts. *Transactions on Machine Learning Research*, 2023.

Rohit Agarwal, Arijit Das, Alexander Horsch, Krishna Agarwal, and Dilip K Prasad. Online learning under haphazard input conditions: A comprehensive review and analysis. *arXiv preprint arXiv:2404.04903*, 2024.

Rohit Agarwal, Aryan Dessai, Arif Ahmed Sekh, Krishna Agarwal, Alexander Horsch, and Dilip K Prasad. Haphazard inputs as images in online learning. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, 2025a.

Rohit Agarwal, Karaka Prasanth Naidu, Alexander Horsch, Krishna Agarwal, and Dilip Prasad. packetLSTM: Dynamic LSTM framework for streaming data with varying feature space, 2025b. URL https://openreview.net/forum?id=VzdycorGTt.

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. *Advances in neural information processing systems*, 32, 2019.

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

Ege Beyazit, Jeevithan Alagurajah, and Xindong Wu. Online learning from data streams with varying feature spaces. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3232–3239, 2019.

Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Philipp Kranen, Hardy Kremer, Timm Jansen, and Thomas Seidl. Moa: Massive online analysis, a framework for stream classification and clustering. In *Proceedings of the first workshop on applications of pattern analysis*, pp. 44–50, 2010.

Albert Bifet, Ricard Gavalda, Geoffrey Holmes, and Bernhard Pfahringer. *Machine learning for data streams: with practical examples in MOA*. MIT press, 2023.

Rudolf K Bock, A Chilingarian, M Gaug, Frantisek Hakl, Thomas Hengstebeck, Marcel Jiřina, Jan Klaschka, Emil Kotrč, Petr Savickỳ, Sherry Towers, et al. Methods for multidimensional event classification: a case study using images from a cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 516(2-3):511–528, 2004.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159, 2021.

Himanshu Buckchash, Momojit Biswas, Rohit Agarwal, and Dilip K Prasad. Hedging is not all you need: A simple baseline for online learning under haphazard inputs. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480, 2015.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

A Philip Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2): 278–290, 1984.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Jessilyn Dunn, Lukasz Kidzinski, Ryan Runge, Daniel Witt, Jennifer L Hicks, Sophia Miryam Schüssler-Fiorenza Rose, Xiao Li, Amir Bahmani, Scott L Delp, Trevor Hastie, et al. Wearable sensors enable personalized predictions of clinical laboratory measurements. *Nature medicine*, 27(6):1105–1112, 2021.

Ryan Elwell and Robi Polikar. Incremental learning of concept drift in nonstationary environments. *IEEE transactions on neural networks*, 22(10):1517–1531, 2011.

Conor Fahy, Shengxiang Yang, and Mario Gongora. Scarcity of labels in non-stationary data streams: A survey. *ACM Computing Surveys (CSUR)*, 55(2):1–39, 2022.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, pp. 23–37, 1995.

Joao Gama, Raquel Sebastiao, and Pedro Pereira Rodrigues. On evaluating stream learning algorithms. *Machine learning*, 90(3):317–346, 2013.

Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D Lee. Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Sanjam Garg, Shafi Goldwasser, and Prashant Nalini Vasudevan. Formalizing data deletion in the context of the right to be forgotten. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 373–402, 2020.

Jean-Gabriel Gaudreault and Paula Branco. A systematic literature review of novelty detection in data streams: Challenges and opportunities. *ACM Computing Surveys*, 56(10):1–37, 2024.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33:16937–16947, 2020.

Peiliang Gong, Mohamed Ragab, Min Wu, Zhenghua Chen, Yongyi Su, Xiaoli Li, and Daoqiang Zhang. Augmented contrastive clustering with uncertainty-aware prototyping for time series test time adaptation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 390–401, 2025.

Stephen Grossberg. How does a brain build a cognitive code? *Psychological review*, 87(1):1, 1980.

Michael Harries, New South Wales, et al. Splice-2 comparative evaluation: Electricity pricing. 1999.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yi He, Jiaxian Dong, Bo-Jian Hou, Yu Wang, and Fei Wang. Online learning in variable feature spaces with mixed data. In *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 181–190, 2021.

Yi He, Christian Schreckenberger, Heiner Stuckenschmidt, and Xindong Wu. Towards utilitarian online learning-a review of online algorithms in open feature space. In *IJCAI*, pp. 6647–6655, 2023.

Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Reeber Erik Forman George Hopkins, Mark and Jaap Suermondt. Spambase. UCI Machine Learning Repository, 1999. DOI: https://doi.org/10.24432/C53G6X.

Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification. 2003.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hao Huang and Scott Evans. Flowfill: A fast and energy-efficient streaming framework for recovering missing values in industrial sensor data. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pp. 4262–4268, 2025.

Jeonghwan Im and Hyuk-Yoon Kwon. Cosa: Context-aware output-space adapter for test-time adaptation in time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026.

Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva. *Measuring attitudes cross-nationally: Lessons from the European Social Survey*. 2007.

Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pp. 202–207, 1996.

Ying-yee Ava Lau, Zhiwen Shao, and Dit-Yan Yeung. Fast and slow streams for online time series forecasting without information leakage. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuecheng Li, Lele Fu, Tong Wang, Jian Lou, Bin Chen, Lei Yang, Jian Shen, Zibin Zheng, and Chuan Chen. Clients collaborate: Flexible differentially private federated learning with guaranteed improvement of utility-privacy trade-off. In *Forty-second International Conference on Machine Learning*, 2025.

Heng Lian, Di Wu, Bo-Jian Hou, Jian Wu, and Yi He. Online learning from evolving feature spaces with deep variational models. *IEEE Transactions on Knowledge and Data Engineering*, 36(8): 4144–4162, 2024.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

Olvi L Mangasarian and William H Wolberg. Cancer diagnosis via linear programming. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 1990.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282, 2017.

H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1222–1230, 2013.

Aditya Krishna Menon, Andreas Veit, Ankit Singh Rawat, Himanshu Jain, Sadeep Jayasumana, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR) 2021*, 2021.

Rami Mohammad and Lee McCluskey. Phishing websites. UCI Machine Learning Repository, 2012. URL `https://doi.org/10.24432/C51W2X`.

Hai-Long Nguyen, Yew-Kwong Woon, and Wee-Keong Ng. A survey on data stream clustering and classification. *Knowledge and information systems*, 45(3):535–569, 2015.

Agustín Alejandro Ortiz Díaz, Sergio Cleger Tamayo, Gabriel Dos Santos Lima, Geovana Amorim Abensur, and Delrick Nunes De Oliveira. Feasibility study to adapt online deep learning models for immersive environments. In *International Conference on Human-Computer Interaction*, pp. 308–322, 2025.

Quang Pham, Chenghao Liu, Doyen Sahoo, and Steven CH Hoi. Learning fast and slow for online time series forecasting. *arXiv preprint arXiv:2202.11672*, 2022.

Quang Pham, Chenghao Liu, and Steven CH Hoi. Continual learning, fast and slow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):134–149, 2024.

Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. Forgetting personal data and revoking consent under the gdpr: Challenges and proposed solutions. *Journal of cybersecurity*, 4(1): tyy001, 2018a.

Eugenia Politou, Alexandra Michota, Efthimios Alepis, Matthias Pocs, and Constantinos Patsakis. Backups and the right to be forgotten in the gdpr: An uneasy relationship. *Computer Law & Security Review*, 34(6):1247–1257, 2018b.

J. Ross Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3): 221–234, 1987.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report, 2025.

Doyen Sahoo, Quang Pham, Jing Lu, and Steven C. H. Hoi. Online deep learning: learning deep neural networks on the fly. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 2660–2666, 2018.

Jeffrey C Schlimmer and Richard H Granger Jr. Incremental learning from noisy data. *Machine learning*, 1(3):317–354, 1986.

Christian Schreckenberger, Yi He, Stefan Lüdtke, Christian Bartelt, and Heiner Stuckenschmidt. Online random feature forests for learning in varying feature spaces. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 4587–4595, 2023.

Ayush Sekhari, Karthik Sridharan, and Satyen Kale. Sgd: The role of implicit regularization, batch-size and multiple-epochs. *Advances In Neural Information Processing Systems*, 34:27422–27433, 2021.

Tegjyot Singh Sethi and Mehmed Kantardzic. On the reliable detection of concept drift from streaming unlabeled data. *Expert Systems with Applications*, 82:77–99, 2017.

Alen David Shapiro. The role of structured induction in expert systems. 1984.

Jack W Smith, James E Everhart, William C Dickson, William C Knowler, and Robert Scott Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, pp. 261, 1988.

W Nick Street and YongSeog Kim. A streaming ensemble algorithm (sea) for large-scale classification. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 377–382, 2001.

W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pp. 861–870, 1993.

Rui Su, Husheng Guo, and Wenjian Wang. Elastic online deep learning for dynamic streaming data. *Information Sciences*, 676:120799, 2024.

Muhammad Sulaiman, Mina Farmanbar, Shingo Kagami, Ahmed Nabil Belbachir, and Chunming Rong. Online deep learning's role in conquering the challenges of streaming data: a survey. *Knowl. Inf. Syst.*, 67(4):3159–3203, 2025.

Haoran Tang, Shiqing Wu, Zhihong Cui, Yicong Li, Guandong Xu, and Qing Li. Model-agnostic dual-side online fairness learning for dynamic recommendation. *IEEE Trans. on Knowl. and Data Eng.*, 37(5):2727–2742, 2025.

Noam Tirosh. Reconsidering the 'right to be forgotten'–memory rights and the right to memory in the new media era. *Media, Culture & Society*, 39(5):644–660, 2017.

Antonios Valkanas, Boris N. Oreshkin, and Mark Coates. Modl: Multilearner online deep learning. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 4321–4329, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Eduard Fosch Villaronga, Peter Kieseberg, and Tiffany Li. Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review*, 34(2): 304–313, 2018.

Heqiang Wang, Xiang Liu, Xiaoxiong Zhong, Lixing Chen, Fangming Liu, and Weizhe Zhang. Multimodal online federated learning with modality missing in internet of things. *IEEE Transactions on Mobile Computing*, 25(2):2172–2185, 2026a.

Yulin Wang, Yi He, Dianlong You, and Di Wu. Online learning for wild streaming data with delayed feedback. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5196–5207, 2026b.

Qingsong Wen, Weiqi Chen, Liang Sun, Zhang Zhang, Liang Wang, Rong Jin, Tieniu Tan, et al. Onenet: Enhancing time series forecasting models under concept drift by online ensembling. *Advances in Neural Information Processing Systems*, 36:69949–69980, 2023.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Luyu Yang, Mingfei Gao, Zeyuan Chen, Ran Xu, Abhinav Shrivastava, and Chetan Ramaiah. Burn after reading: Online adaptation for cross-domain streaming data. In *European Conference on Computer Vision*, pp. 404–422, 2022.

Dianlong You, Huigui Yan, Jiawei Xiao, Zhen Chen, Di Wu, Limin Shen, and Xindong Wu. Online learning for data streams with incomplete features and labels. *IEEE Transactions on Knowledge and Data Engineering*, 36(9):4820–4834, 2024.

Linglan Zhao, Xuerui Zhang, Ke Yan, Shouhong Ding, and Weiran Huang. Safe: Slow and fast parameter-efficient tuning for continual learning with pre-trained models. *Advances in Neural Information Processing Systems*, 37:113772–113796, 2024.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

Shengda Zhuo, Di Wu, Yi He, Shuqiang Huang, and Xindong Wu. Online learning from mix-typed, drifted, and incomplete streaming features. *ACM Transactions on Knowledge Discovery from Data*, 2025.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936, 2003.

Indrė Žliobaitė. Combining similarity in time and space for training set formation under concept drift. *Intelligent Data Analysis*, 15(4):589–611, 2011.

TABLE OF CONTENTS

A Supplementary Background on Data Retention and Privacy **17**

B Supplementary Related Work **17**

C Supplementary Method Details **18**

C.1 Mask-Aware Causal Normalization 18

C.2 OnlineNet Pseudocode 18

D Supplementary Experimental Settings **18**

D.1 Supplementary Datasets Description 18

D.2 Supplementary Metrics Description 22

D.3 Supplementary Baseline Descriptions 23

D.4 Supplementary Implementation Details 23

E Supplementary Experimental Results and Analysis **26**

E.1 Supplementary Results on Cumulative Performance 26

E.2 Supplementary Results on Online Performance Evolution 34

E.3 Supplementary Complexity Analysis 34

E.4 Supplementary Ablation Studies 36

E.5 Supplementary Hyperparameter Analysis 36

E.5.1 Supplementary Results on Modulation Coefficient 36

E.5.2 Supplementary Results on Adjustment Coefficient 36

F Additional Exploratory Analyses **38**

F.1 Learning-Rate Sensitivity of the Fast and Slow Pathways 38

F.2 Exploring Transformer-Based Online Architectures 38

A Supplementary Background on Data Retention and Privacy

Privacy-Sensitive Streaming Application Streaming applications frequently process information closely associated with individual users. Depending on the application, continuously arriving observations may encode behavioral information such as user interactions and preferences Tang et al. (2025), contextual information collected from mobile environments Wang et al. (2026a), or physiological measurements obtained from health-monitoring devices Dunn et al. (2021). Unlike static datasets collected for one-time analysis, streaming systems continuously accumulate such observations over time. Persistently retaining them can therefore create an increasingly extensive record of individual activity and enlarge the exposure surface for sensitive information. Moreover, observations collected for immediate prediction or adaptation may remain available for subsequent reuse long after their original purpose has been fulfilled, raising additional concerns Yang et al. (2022) regarding data minimization, consent management, and long-term data governance.

Persistent Data Retention and the Right to Erasure The retention of streaming data also creates practical challenges for subsequent data erasure. The Right to Be Forgotten (RTBF) Tirosh (2017); Garg et al. (2020), formally expressed as the right to erasure under Article 17 of the General Data Protection Regulation (GDPR), provides data subjects with the right to obtain the erasure of their personal data under specified conditions.[1] In particular, Article 17 identifies cases in which personal data are no longer necessary for the purposes for which they were collected or otherwise processed. In streaming systems, however, continuously generated observations may be retained across operational databases, backup systems, and learning pipelines, making subsequent erasure increasingly difficult once historical records have been persistently stored and reused Politou et al. (2018a;b); Villaronga et al. (2018). This challenge is also reflected in recent regulatory evidence. In its 2025 Coordinated Enforcement Action on the implementation of the right to erasure, 32 supervisory authorities across the EEA participated, with responses collected from 764 controllers. The resulting report identified seven recurring issues, including difficulties in defining and implementing data-retention periods and deleting personal data from backup systems.[2] These findings provide practical evidence of the difficulties associated with data retention and subsequent erasure even when deletion rights are formally established. Accordingly, this work adopts a *use-and-discard* principle for streaming learning, aiming to reduce unnecessary persistent data retention.

B Supplementary Related Work

Privacy-Oriented Learning Paradigms Several learning paradigms partially alleviate privacy concerns but address different stages of the data lifecycle. Federated learning McMahan et al. (2017); Li et al. (2025) avoids centralizing raw data by keeping training observations on local devices. However, it does not eliminate the need to retain these observations locally, and transmitted model updates may still reveal information about the underlying data Zhu et al. (2019); Geiping et al. (2020). In contrast, machine unlearning Cao & Yang (2015); Bourtoule et al. (2021) aims to remove the influence of previously used samples from an already trained model. Such removal is retrospective and may introduce additional computational overhead or require access to historical information. These paradigms therefore address privacy through different mechanisms and objectives from those considered in this paper and fall outside the scope of our study.

[1] Article 17 GDPR: Right to erasure (right to be forgotten). https://gdpr.eu/article-17-right-to-be-forgotten/

[2] European Data Protection Board, *2025 Coordinated Enforcement Action: Implementation of the right to erasure by controllers*, adopted 10 February 2026. https://www.edpb.europa.eu/system/files/2026-02/edpb_cef-report_2025_right-to-erasure_en.pdf

C Supplementary Method Details

C.1 Mask-Aware Causal Normalization

OnlineNet maintains feature-wise running statistics without retaining historical observations. For feature j, let $(n_{t-1,j}, \mu_{t-1,j}, M_{2,t-1,j})$ denote the Welford statistics available before round t, where $n_{t-1,j}$ counts previously observed values, $\mu_{t-1,j}$ is their running mean, and $M_{2,t-1,j}$ accumulates the corresponding squared deviations. The variance and numerical scale used for the current normalization are computed as:

$$v_{t-1,j} = \begin{cases} M_{2,t-1,j}/(n_{t-1,j}-1), & n_{t-1,j} > 1, \\ 1, & n_{t-1,j} \le 1, \end{cases} \tag{10}$$

$$q_{t-1,j} = \begin{cases} \sqrt{v_{t-1,j}}, & \sqrt{v_{t-1,j}} > \epsilon_x, \\ 1, & \text{otherwise}, \end{cases} \qquad \epsilon_x = 10^{-8}. \tag{11}$$

Using only statistics available before round t, the normalized value of feature j is obtained as:

$$z_{t,j} = m_{t,j} \frac{\widetilde{x}_{t,j} - \mu_{t-1,j}}{q_{t-1,j}}. \tag{12}$$

The multiplication by $m_{t,j}$ keeps unavailable features at zero, while the statistics used for normalization are determined only by previously observed values.

Only after the prediction and the supervised model update are completed is the current observation incorporated into the running statistics. For each observed feature j, the mask-aware Welford update is:

$$\xi_{t,j} = \widetilde{x}_{t,j} - \mu_{t-1,j}, \tag{13}$$

$$n_{t,j} = n_{t-1,j} + m_{t,j}, \tag{14}$$

$$\mu_{t,j} = \mu_{t-1,j} + \frac{m_{t,j}\xi_{t,j}}{\max(n_{t,j}, 1)}, \tag{15}$$

$$M_{2,t,j} = M_{2,t-1,j} + m_{t,j}\xi_{t,j}\left(\widetilde{x}_{t,j} - \mu_{t,j}\right). \tag{16}$$

When $m_{t,j} = 0$, the corresponding running statistics remain unchanged. Since Equation 12 uses the statistics available before round t and Equation 16 is applied only after the current supervised update, the current observation cannot influence its own normalization. This ordering implements the causal normalization used by the slow pathway without storing historical observations.

C.2 OnlineNet Pseudocode

Algorithm 1 summarizes the test-then-train procedure of OnlineNet. All predictive quantities at round t are computed from the current input and the states available before y_t is revealed. The causal stream state and class counts are updated only after the supervised model update.

D Supplementary Experimental Settings

D.1 Supplementary Datasets Description

For clarity, we group the datasets by stream length into small-scale streams ($n < 10{,}000$) and large-scale streams ($n \ge 10{,}000$). Dataset statistics and scale assignments are reported in Table 4, with dataset-specific details provided below.

Real-world Datasets We use 15 publicly available real-world classification datasets drawn from the UCI Machine Learning Repository Asuncion et al. (2007), the LIBSVM dataset repository Chang & Lin (2011), and established concept-drift benchmarks Žliobaitė (2011); Sethi & Kantardzic (2017); Elwell & Polikar (2011); Harries et al. (1999). Among these datasets, LUdata, spam, weather, and elec preserve temporal order and exhibit naturally occurring concept drift, although the corresponding change points and drift types are not annotated. The specific description is as follows:

Algorithm 1: OnlineNet

Input : Stream $\{(\mathbf{x}_t, \mathbf{m}_t, y_t)\}_{t=1}^{T}$; initial parameters Θ_1; number of paired fast–slow layers L; modulation coefficient λ; adjustment coefficient τ
Output: Predictions $\{\hat{y}_t\}_{t=1}^{T}$ and updated parameters Θ_{T+1}

Initialize causal stream state $\mathcal{C}_0$ and class counts $\mathbf{N}_0 \leftarrow \mathbf{0}$;
for $t \leftarrow 1$ **to** T **do**
  // Stage 1: Prediction with the pre-update state
  $\widetilde{\mathbf{x}}_t \leftarrow \mathbf{m}_t \odot \mathbf{x}_t$;
  $\mathbf{z}_t \leftarrow \mathrm{CausalNormalize}(\widetilde{\mathbf{x}}_t, \mathbf{m}_t, \mathcal{C}_{t-1})$;
  $\mathbf{f}_{0,t} \leftarrow [\widetilde{\mathbf{x}}_t; \mathbf{m}_t]$ and $\mathbf{h}_{0,t} \leftarrow \mathbf{z}_t$;
  for $\ell \leftarrow 1$ **to** L **do**
    $\mathbf{f}_{\ell,t} \leftarrow \mathrm{ReLU}(\mathbf{W}_\ell^f \mathbf{f}_{\ell-1,t} + \mathbf{b}_\ell^f)$;
    $\mathbf{s}_{\ell,t} \leftarrow \mathrm{ReLU}(\mathbf{W}_\ell^s \mathbf{h}_{\ell-1,t} + \mathbf{b}_\ell^s)$;
    $\mathbf{g}_{\ell,t} \leftarrow \mathbf{1} + \lambda(2\sigma(\mathbf{f}_{\ell,t}) - \mathbf{1})$;
    $\mathbf{h}_{\ell,t} \leftarrow \mathbf{s}_{\ell,t} \odot \mathbf{g}_{\ell,t}$;
  $\mathbf{o}_t \leftarrow \mathbf{W}^{\mathrm{out}} \mathbf{h}_{L,t} + \mathbf{b}^{\mathrm{out}}$;
  $\hat{y}_t \leftarrow \arg\max_{c \in \mathcal{Y}} o_{t,c}$;
  $\widehat{\boldsymbol{\pi}}_t \leftarrow \mathrm{PrequentialPrior}(\mathbf{N}_{t-1}, \alpha)$;
  Emit $\hat{y}_t$ and observe y_t;
  // Stage 2: Single-instance model update
  $\overline{\mathbf{o}}_t \leftarrow \mathbf{o}_t + \tau \log \widehat{\boldsymbol{\pi}}_t$;
  $\mathcal{L}_t \leftarrow -\log \mathrm{softmax}(\overline{\mathbf{o}}_t)_{y_t}$;
  $\Theta_{t+1} \leftarrow \mathrm{Update}(\Theta_t, \mathcal{L}_t)$;
  // Stage 3: Causal state update
  $\mathcal{C}_t \leftarrow \mathrm{UpdateFeatureState}(\mathcal{C}_{t-1}, \mathbf{x}_t, \mathbf{m}_t)$;
  $\mathbf{N}_t \leftarrow \mathrm{UpdateClassCounts}(\mathbf{N}_{t-1}, y_t)$;
  Discard $(\mathbf{x}_t, \mathbf{m}_t, y_t)$;

- **wdbc:** The Wisconsin Diagnostic Breast Cancer dataset Street et al. (1993) contains 569 samples represented by 30 numerical features computed from digitized images of fine-needle aspirates of breast masses. The task is to distinguish malignant from benign cases.
- **australian:** The Statlog Australian Credit Approval dataset Quinlan (1987) contains 690 anonymized credit applications with six numerical and eight categorical attributes. The target indicates whether an application is approved.
- **wbc:** The Breast Cancer Wisconsin (Original) dataset Mangasarian & Wolberg (1990) contains 699 fine-needle aspirate samples described by nine numerical cytological features. Each feature is recorded as an ordinal integer score from 1 to 10, and the target distinguishes benign from malignant tumors.
- **diabetes_f:** The Pima Indians Diabetes dataset Smith et al. (1988) comprises 768 clinical records from women of Pima Indian heritage. Each record contains eight numerical health and demographic features used to predict diabetes status.
- **german:** The German Credit dataset Hofmann (1994) contains 1,000 credit records described by 20 mixed numerical/categorical attributes. We use the numerical Statlog representation, which encodes these attributes into 24 numerical input features.
- **svmguide3:** The svmguide3 dataset Hsu et al. (2003) is a binary vehicle-classification benchmark distributed with the LIBSVM dataset collection Chang & Lin (2011). We use its training set, which contains 1,243 instances and 21 numerical features.
- **LUdata:** The Luxembourg dataset (LUdata) Žliobaitė (2011) is constructed from three rounds of the European Social Survey Jowell et al. (2007) collected over five years. Its 1,901 respondents are represented by 31 numerical features, and the task distinguishes high from low internet usage. We retain the survey collection order, while the change points and drift type remain unannotated.
- **kr-vs-kp:** The Chess (King-Rook vs. King-Pawn) dataset Shapiro (1984) contains 3,196 endgame positions represented by 36 categorical features. Each position describes a con-

Table 4: Dataset statistics used in our experiments. The imbalance ratio is measured as the majority-class percentage. Data Type indicates whether the features are numerical (Num.), categorical (Cat.), mixed numerical/categorical (Mix), or binary-valued (Bin.). Drift Type describes the concept-drift setting. The symbol ∘ indicates that neither the drift type nor the change points are annotated, while Abrupt and Gradual denote explicitly controlled abrupt and gradual drift, respectively.

Dataset	Instances	Features	Imbalance	Data Type	Drift Type	Domain
Small-scale streams ($n < 10{,}000$)						
Real-world Datasets						
wdbc	569	30	62.74%	Num.	∘	Medical diagnosis
australian	690	14	55.51%	Mix	∘	Credit assessment
wbc	699	9	65.52%	Num.	∘	Medical diagnosis
diabetes_f	768	8	65.10%	Num.	∘	Medical diagnosis
german	1,000	24	70.00%	Num.	∘	Credit assessment
svmguide3	1,243	21	76.19%	Num.	∘	Vehicle classification
LUdata	1,901	31	51.39%	Num.	∘	Medical/social survey
kr-vs-kp	3,196	36	52.22%	Cat.	∘	Chess endgame
spambase	4,601	57	60.60%	Num.	∘	Spam detection
spam	6,213	499	66.70%	Bin.	∘	Spam detection
Synthetic Datasets						
LED_s-A	5,000	24	50.32%	Bin.	Abrupt	-
LED_s-G	5,000	24	50.32%	Bin.	Gradual	-
SEA_s-A	5,000	3	60.28%	Num.	Abrupt	-
SEA_s-G	5,000	3	60.28%	Num.	Gradual	-
STAGGER_s-A	5,000	3	67.04%	Cat.	Abrupt	-
STAGGER_s-G	5,000	3	67.04%	Cat.	Gradual	-
Large-scale streams ($n \geq 10{,}000$)						
Real-world Datasets						
phishing	11,055	46	55.69%	Bin.	∘	Phishing detection
weather	18,159	8	68.62%	Num.	∘	Weather prediction
magic04	19,020	10	64.84%	Num.	∘	Particle classification
a8a	32,561	123	75.92%	Bin.	∘	Income prediction
elec	45,312	6	57.55%	Num.	∘	Electricity market
Synthetic Datasets						
LED_m-A	50,000	24	50.76%	Bin.	Abrupt	-
LED_m-G	50,000	24	50.73%	Bin.	Gradual	-
SEA_m-A	50,000	3	60.47%	Num.	Abrupt	-
SEA_m-G	50,000	3	60.47%	Num.	Gradual	-
STAGGER_m-A	50,000	3	59.93%	Cat.	Abrupt	-
STAGGER_m-G	50,000	3	59.85%	Cat.	Gradual	-

figuration in which White has a king and rook, Black has a king and pawn, and White is to move. The target indicates whether White can force a win.

- **spambase:** The Spambase dataset Hopkins & Suermondt (1999) contains 4,601 email messages represented by 57 numerical features. The features measure word and character frequencies together with statistics of consecutive capital-letter sequences. Each message is labeled as spam or non-spam.
- **spam:** The Spam dataset Sethi & Kantardzic (2017) forms a temporally ordered stream of 6,213 email messages. Each message is represented by 499 binary-valued word-occurrence features. We retain the supplied order, while the locations and types of changes in spam characteristics remain unannotated.

- **phishing:** The Phishing Websites dataset Mohammad & McCluskey (2012) contains 11,055 websites labeled as phishing or legitimate. The original release describes each website using 30 attributes derived from its URL, webpage, and domain properties. We use the 46-feature binary representation adopted in concept-drift research Sethi & Kantardzic (2017) and retain its supplied order. This processed version provides neither timestamps nor annotated change points.
- **weather:** The Nebraska Weather dataset Elwell & Polikar (2011) contains 18,159 daily observations recorded at Offutt Air Force Base in Bellevue, Nebraska, between 1949 and 1999. Each observation contains eight numerical meteorological features, and the target distinguishes rainy from non-rainy days. The chronological order is preserved, but change points are not annotated.
- **magic04:** The MAGIC Gamma Telescope dataset Bock et al. (2004) contains 19,020 simulated particle-shower events observed by an atmospheric Cherenkov telescope. Ten numerical parameters extracted from each shower image are used to distinguish gamma-ray signals from hadronic background events.
- **a8a:** The a8a dataset distributed by LIBSVM Chang & Lin (2011) is derived from the Adult census-income benchmark Kohavi et al. (1996). We combine its training and test sets, obtaining 32,561 instances in the existing 123-dimensional binary LIBSVM representation.
- **elec:** The Electricity Market dataset Harries et al. (1999) contains 45,312 half-hourly observations from the New South Wales electricity market between May 1996 and December 1998. We use the normalized `elecNormNew` representation, in which the date and weekday fields are omitted, leaving six numerical features. The label indicates whether the New South Wales price is above or below its moving average over the preceding 24 hours. We preserve chronological order, while change points remain unannotated.

Synthetic Datasets To evaluate adaptation under controlled concept changes, we supplement the real-world datasets with synthetic datasets. The streams are generated with MOA 2024.07.0 Bifet et al. (2010) using the LEDLoh (2011), SEA Street & Kim (2001), and STAGGER generators Schlimmer & Granger Jr (1986). For each generator, f_j denotes the j-th predefined classification function. Note that STAGGER can provide a maximum of three predefined concepts. We consider two stream scales. Each small-scale stream contains 5,000 instances and follows $f_1 \to f_2$, with the transition centered at instance 2,500. Each large-scale LED and SEA stream contains 50,000 instances and follows $f_1 \to f_2 \to f_3 \to f_4$, with transitions centered at instances 15,000, 25,000, and 35,000. The large-scale STAGGER streams follow $f_1 \to f_2 \to f_3$, with transitions centered at instances 20,000 and 40,000. We set the transition width to $w = 1$ to approximate abrupt drift. Gradual drift uses $w = 500$ in small-scale streams and $w = 5{,}000$ in large-scale streams.

We implement each transition using MOA's `ConceptDriftStream`. Let $\mathcal{D}_{f_j}$ denote the joint distribution of $(\mathbf{x}, y)$ generated under f_j. For a transition from f_a to f_b, an instance at time t is sampled according to:

$$(\mathbf{x}_t, y_t) \sim (1 - \pi_t)\mathcal{D}_{f_a} + \pi_t \mathcal{D}_{f_b}, \qquad \pi_t = \frac{1}{1 + \exp[-4(t - c)/w]},$$

where c is the configured transition position. At $t = c$, the two concepts have equal sampling probability. The probability of sampling from the new concept increases from approximately 0.12 to 0.88 over $[c - w/2, c + w/2]$. For streams with multiple changes, we nest this construction in the concept order specified above.

To support reproducibility, we vary only the classification functions, stream lengths, transition positions, and widths reported above. All remaining MOA options retain their default values. In particular, the generators and drift wrappers use a random seed of 1, and class balancing is disabled. Each synthetic data is generated once and reused across all methods. Next, we describe the attribute distributions and classification functions used by the three generators:

- **LED:** The LED generator produces 24 binary attributes and assigns one of ten digit labels. We use the first four classification functions implemented by MOA. The observed attributes are binary display and distractor bits, and the raw label under concept f_j is:

$$y^{\text{raw}} = f_j(\mathbf{x}), \qquad y^{\text{raw}} \in \{0, \ldots, 9\}.$$

To make the generated streams consistent with the binary classification setting used in our experiments, we map the raw digit label according to $y = \lfloor y^{\text{raw}}/5 \rfloor$, assigning digits 0–4 to class 0 and digits 5–9 to class 1. This conversion is applied after stream generation and before online training and evaluation, without altering the attributes or their temporal order.

- **SEA:** The SEA generator independently samples three numerical attributes according to:

$$x_1, x_2, x_3 \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, 10).$$

Only x_1 and x_2 determine the clean label. Under concept f_j, the clean label is defined as:

$$y^\star = \mathbb{I}[x_1 + x_2 > \theta_j], \qquad (\theta_1, \theta_2, \theta_3, \theta_4) = (8, 9, 7, 9.5).$$

A concept change shifts the decision boundary without changing the attribute distribution. We apply MOA's default 10% label-noise rate as follows:

$$y = y^\star \oplus b, \qquad b \sim \text{Bernoulli}(0.1),$$

where $\oplus$ denotes binary exclusive OR.

- **STAGGER:** The STAGGER generator independently and uniformly samples size, color, and shape from their respective sets of three categorical values. The positive class under its three concepts is defined by:

$$f_1(\mathbf{x}) = \mathbb{I}[\text{size} = \text{small} \wedge \text{color} = \text{red}],$$
$$f_2(\mathbf{x}) = \mathbb{I}[\text{color} = \text{green} \vee \text{shape} = \text{circle}],$$
$$f_3(\mathbf{x}) = \mathbb{I}[\text{size} \in \{\text{medium}, \text{large}\}].$$

D.2 Supplementary Metrics Description

We evaluate classification performance using five metrics: Accuracy, AUROC, AUPRC, Balanced Accuracy, and Sliding-Window Error. The first four are cumulative metrics computed over the stream. Sliding-Window Error characterizes local predictive performance over a recent window.

Accuracy The fraction of correctly classified instances is computed as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{17}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Balanced Accuracy The average of sensitivity and specificity is computed as:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right). \tag{18}$$

This metric assigns equal importance to both classes and is therefore more robust under class imbalance.

AUROC The area under the receiver operating characteristic (ROC) curve evaluates the ranking quality of predicted scores across positive and negative instances. It is threshold-independent and can be interpreted as the probability that a randomly selected positive instance receives a higher prediction score than a randomly selected negative instance.

AUPRC The area under the precision-recall curve summarizes the trade-off between precision and recall, particularly under imbalanced class distributions.

Sliding-Window Error The sliding-window error measures the classification error rate over recent instances. Given a maximum window length W, it is computed as:

$$\text{Error}_t = \frac{1}{\min(t, W)} \sum_{i=\max(1, t-W+1)}^{t} \mathbb{I}(\hat{y}_i \neq y_i), \tag{19}$$

where $\mathbb{I}(\cdot)$ is the indicator function, y_i denotes the ground-truth label, and $\hat{y}_i$ denotes the predicted label. For $t < W$, the metric is computed using all available predictions.

D.3 Supplementary Baseline Descriptions

We also considered several related online learning paradigms that may appear relevant but follow learning protocols different from the setting studied in this work. Online time-series forecasting Pham et al. (2022); Wen et al. (2023); Lau et al. (2025) typically predicts future observations from a temporal context consisting of historical sequences, and therefore requires repeated access to past observations as model inputs. Online continual learning Aljundi et al. (2019); Pham et al. (2024) primarily studies knowledge acquisition and forgetting across evolving tasks, with assuming explicit or implicit task boundaries or other task-level structure. In contrast, test-time adaptation Gong et al. (2025); Im & Kwon (2026), adapts a pretrained model to distribution changes during inference without access to ground-truth labels, typically relying on unsupervised or self-supervised adaptation objectives. Since these paradigms differ from the online classification setting considered in this work, we do not include them as empirical baselines.

We further examined recently proposed online learning methods that are more closely related to our setting, including EODL Su et al. (2024), PacketLSTM Agarwal et al. (2025b), and HapNet Buckchash et al. (2025). However, these methods are not included in the empirical comparison because their implementations are not publicly available or the publicly available implementation details are insufficient for faithful reproduction. We therefore compare OnlineNet with eight representative and reproducible baselines spanning classical online learning and online deep learning, as summarized in Table 5.

D.4 Supplementary Implementation Details

All models are implemented in PyTorch. Experiments are conducted on a machine equipped with a 13th Gen Intel(R) Core(TM) i7-13700 CPU at 2.10 GHz. Since online learning processes streaming instances sequentially, CPU execution is used for the serialized data-processing pipeline.

Baselines For ORF^{3}V, OLVF, and Aux-Drop, we use the implementations and settings described in Agarwal et al. (2024). For HBP Sahoo et al. (2018), OLIFL You et al. (2024), HI^2 Agarwal et al. (2025a), and MODL Valkanas et al. (2025), we follow their original papers. Following the configuration of HI^2, we use an ImageNet-pretrained ResNet-34 He et al. (2016) backbone. Aux-Drop Agarwal et al. (2023) assumes at least one consistently available base feature. To apply it in our no-base-feature setting, we follow the adaptation in Agarwal et al. (2024) and designate the first hidden layer as the auxiliary layer. OLWDF is included as a storage-based reference. We set its buffer size to five to enable normal operation under our experimental protocol. OLWDF Wang et al. (2026b) was originally designed for streaming settings with delayed feedback. In our protocol, each label is revealed immediately after prediction, so its delay-handling mechanism is inactive. Therefore, its update procedure reduces to that of OVFM He et al. (2021). We use the OVFM implementation and refer to it as OLWDF throughout the experiments. Figure 5 shows that larger buffers do not consistently improve predictive performance on the evaluated datasets. Larger buffers also increase the amount of retained raw data, which may raise additional privacy concerns. The remaining methods update on the currently arriving instance without retaining past raw instances for reuse.

Our Method OnlineNet uses two paired fast–slow layers with a hidden dimension of 64. The BAM modulation coefficient in Equation 4 is set to $\lambda = 1$. The OPA adjustment coefficient in Equation 8 is set to $\tau = 0.25$. Model parameters are updated using AdamW with weight decay set to 0.01. The learning rate is 10^{-3} for most datasets, 2×10^{-3} for wbc and spambase, and 5×10^{-4} for a8a and spam. The default network depth is two paired layers. We use one layer for a8a.

Table 5: Detailed comparison of the evaluated methods. Methods marked by † and ‡ are classical and deep online learners, respectively. "Buffer" denotes retention of prior instances. "Ext. model" denotes use of an externally pretrained model. "Struct. prior" denotes a predefined partition or relation among input features.

Method	Buffer	Ext. model	Struct. prior	Description
ORF^{3}V†	×	×	×	Each observed feature is modeled by a randomized decision-stump forest, and forest predictions are combined with online weights. Class-conditional t-digests support updates without retaining instances, while a Hoeffding-bound test prunes vanished-feature forests.
OLIFL†	×	×	×	A global feature space and recursively updated informativeness matrix weight observed features. The linear classifier uses available labels and pseudo-labels retained by an error-bound confidence test.
OLVF†	×	×	×	An instance classifier predicts in the shared feature subspace, while a feature-space classifier estimates projection confidence. These estimates scale the soft-margin update before low-weight features are truncated.
OLWDF†	✓	×	×	Coarse and fine buffers support an online Gaussian-copula latent space and a dynamic geometric graph. Calibrated pseudo-labels drive a transient learner, which is fused with a persistent learner trained on delayed labels.
HBP‡	×	×	×	Layer-wise classifiers produce predictions at different depths and are combined with Hedge weights updated from the current losses. The Hedge-weighted layer losses are backpropagated through the shared network.
Aux-Drop‡	×	×	✓	Auxiliary features enter at a designated layer, where unavailable nodes are dropped selectively and the remaining nodes undergo random dropout. The original formulation requires a nonempty base-feature set.
MODL‡	×	×	×	Recursive Bayesian logistic regression, an MLP, and a set learner are jointly trained in a stacked architecture. Their latent class scores are summed before a shared softmax.
HI^2‡	×	✓	×	Observed features are normalized online and rendered as a fixed-size bar-chart image using persistent feature colors. An ImageNet-pretrained vision backbone is updated after each labeled instance.
OnlineNet‡	×	×	×	Causal stream context and instantaneous evidence are encoded through slow and fast pathways, coordinated by BAM, and trained with OPA.

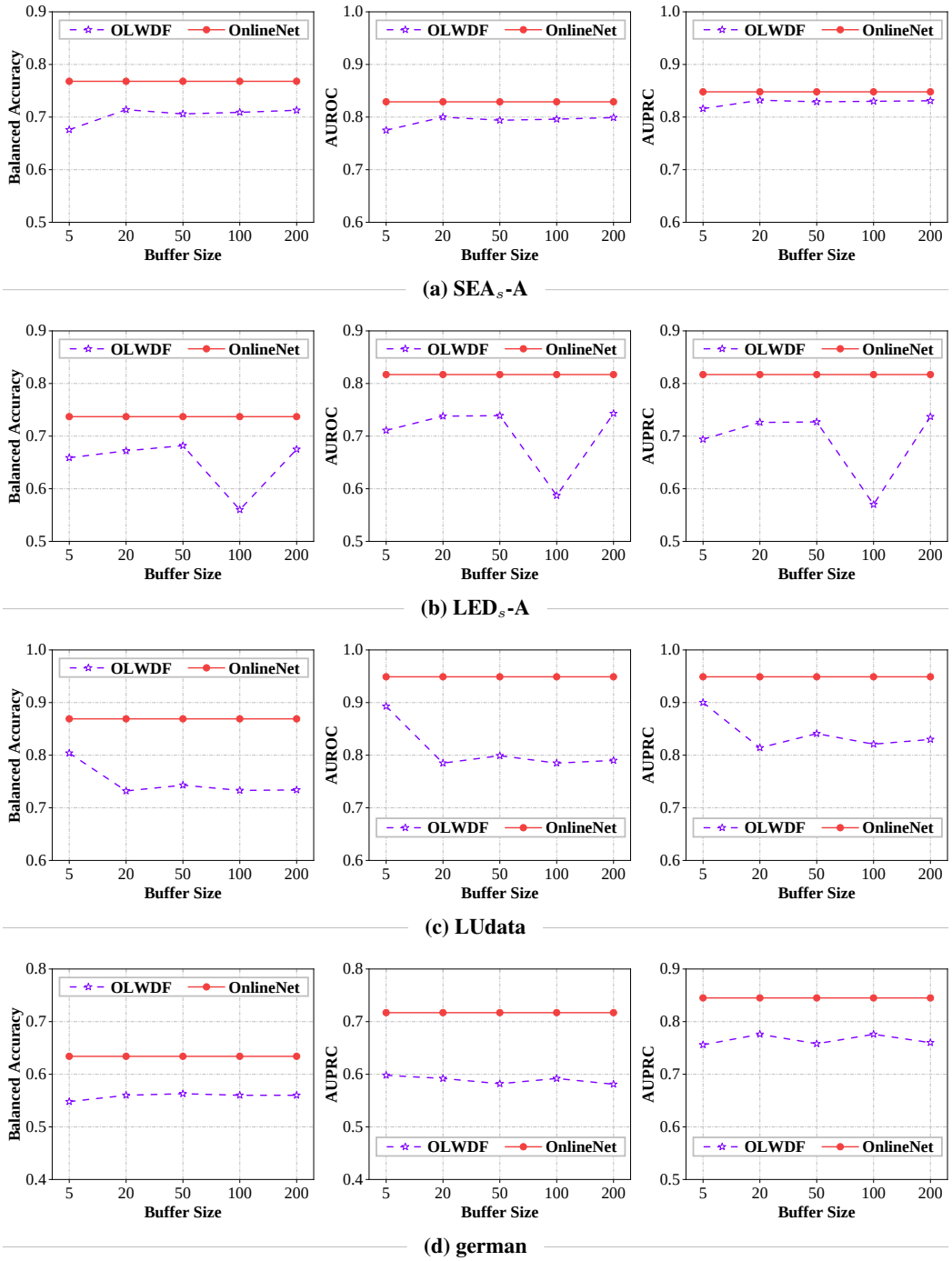


Figure 5: Effect of buffer size on OLWDF performance at $p = 0.75$. Each row corresponds to the dataset named below it. Columns show balanced accuracy, AUROC, and AUPRC, respectively. OnlineNet, which does not retain past instances, is included for comparison.

Table 6: Statistical results over all 81 streaming conditions. Avg. Rank is computed among the nine methods and lower values indicate better performance. Win/Loss is counted from the row method's perspective against OnlineNet. The OnlineNet row aggregates comparisons against all eight baselines. p_{adj} denotes the Holm-adjusted p-value.

Metric	Method	Avg. Rank ↓	Win/Loss	p_{adj}
Accuracy	ORF^{3}V	8.22	1/80	2.44×10^{-4}
	OLIFL	7.09	0/81	2.44×10^{-4}
	OLVF	6.52	0/81	2.44×10^{-4}
	OLWDF	3.69	2/79	2.44×10^{-4}
	HBP	6.25	0/81	2.44×10^{-4}
	Aux-Drop	5.94	0/81	2.44×10^{-4}
	MODL	3.72	0/81	2.44×10^{-4}
	HI^2	2.37	13/68	2.44×10^{-4}
	OnlineNet	**1.20**	632/16	–
AUROC	ORF^{3}V	8.06	0/81	2.44×10^{-4}
	OLIFL	8.35	0/81	2.44×10^{-4}
	OLVF	6.04	0/81	2.44×10^{-4}
	OLWDF	3.34	0/81	2.44×10^{-4}
	HBP	6.15	0/81	2.44×10^{-4}
	Aux-Drop	5.81	0/81	2.44×10^{-4}
	MODL	3.91	0/81	2.44×10^{-4}
	HI^2	2.33	1/80	2.44×10^{-4}
	OnlineNet	**1.01**	647/1	–
AUPRC	ORF^{3}V	8.54	0/81	2.44×10^{-4}
	OLIFL	6.05	6/75	2.44×10^{-4}
	OLVF	6.60	0/81	2.44×10^{-4}
	OLWDF	3.30	0/81	2.44×10^{-4}
	HBP	6.50	0/81	2.44×10^{-4}
	Aux-Drop	6.19	0/81	2.44×10^{-4}
	MODL	4.26	0/81	2.44×10^{-4}
	HI^2	2.48	1/80	2.44×10^{-4}
	OnlineNet	**1.09**	641/7	–
Balanced Accuracy	ORF^{3}V	8.78	0/81	2.44×10^{-4}
	OLIFL	6.09	1/80	2.44×10^{-4}
	OLVF	5.99	0/81	2.44×10^{-4}
	OLWDF	3.59	2/79	2.44×10^{-4}
	HBP	6.65	0/81	2.44×10^{-4}
	Aux-Drop	6.03	0/81	2.44×10^{-4}
	MODL	4.14	0/81	2.44×10^{-4}
	HI^2	2.67	2/79	2.44×10^{-4}
	OnlineNet	**1.06**	643/5	–

E Supplementary Experimental Results and Analysis

E.1 Supplementary Results on Cumulative Performance

Statistical Results Table 6 summarizes the statistical comparisons between OnlineNet and the eight baselines over all 81 streaming conditions. Pairwise comparisons use the Wilcoxon signed-rank statistic Wilcoxon (1945) with exact sign randomization over 18 dependency clusters, comprising one cluster for each real-world dataset and one for each synthetic generator family. Holm correction Holm (1979) is applied jointly across all 32 comparisons. OnlineNet achieves the best average rank across all four metrics and significantly outperforms every baseline after correction ($p_{\mathrm{adj}} < 0.05$). Together with the Win/Loss results, these findings indicate that OnlineNet's advantage is broadly distributed across evaluation metrics and experimental settings rather than driven by a limited number of favorable cases.

Dataset-level Cumulative Performance Tables 7–12 report the complete dataset-level results underlying the aggregate comparison in Table 1. The six tables separately present small- and large-scale streams under $p \in \{0.75, 0.50, 0.25\}$, allowing the aggregate trends to be examined across individual datasets and different levels of feature availability. The detailed results are broadly consistent with the findings in the main text. OnlineNet remains highly competitive across dataset scales, availability ratios, and evaluation metrics, while HI^2 is generally the strongest competing baseline. Large gains occur on several real-world streams. For example, on kr-vs-kp, OnlineNet achieves absolute AUPRC gains of 10.16 percentage points at $p = 0.50$ (84.35% versus 74.19% for HI^2) and 8.76 percentage points at $p = 0.25$ (71.48% versus 62.72% for OLWDF). At $p = 0.25$ on spambase, its absolute gains over the strongest baseline are 7.52 percentage points in AUROC, 8.12 percentage points in AUPRC, and 7.57 percentage points in balanced accuracy. A similar advantage is observed on the large-scale elec stream, where OnlineNet achieves absolute gains of 3.28 percentage points and 5.15 percentage points in AUROC and AUPRC, respectively, at $p = 0.25$. These examples show that the advantage extends to real-world streams of different scales and becomes particularly clear under severe feature unavailability. The best-performing method still varies across individual datasets, and no single method consistently ranks first across all streams. Nevertheless, OnlineNet maintains the most consistent overall performance across datasets, feature observation probabilities, and evaluation metrics, in line with the aggregate findings in the main text.

Strong Performance of HI^2 HI^2 outperforms the main competing baselines on most datasets. Its pretrained vision backbone and fixed-dimensional input transformation may contribute to this performance by providing an external representation prior and reducing the direct effect of missing and varying features. Unlike buffer-dependent OLWDF, HI^2 also does not depend on retaining recent observations for prediction. However, these design choices incur substantially higher per-instance runtime than most competing methods, as shown in Figure 2. HI^2 therefore exhibits a less favorable runtime profile in the evaluated settings.

Table 7: Detailed results on small-scale streams ($n < 10{,}000$) at $p = 0.75$. Higher is better for all metrics. Bold and underlined values indicate the best and second-best results, respectively.

Dataset	Metric	Classical Online Learning				Online Deep Learning				
		OLIFL	OLVF	OLWDF	ORF^{3}V	HBP	Aux-Drop	MODL	HI^2	Our
Real-world datasets										
wdbc	Acc.	58.52	58.35	59.23 (±12.50)	44.96 (±1.01)	65.20 (±1.55)	62.07 (±1.80)	62.50 (±0.62)	87.87 (±1.60)	**92.72** (±1.90)
	AUROC	49.10	68.52	52.31 (±16.46)	38.53 (±1.15)	60.82 (±1.68)	48.30 (±6.78)	50.26 (±0.60)	95.07 (±0.46)	**97.24** (±0.71)
	AUPRC	61.71	66.06	49.48 (±13.29)	31.86 (±0.57)	51.60 (±2.92)	37.21 (±3.24)	39.65 (±1.75)	93.48 (±0.74)	**96.48** (±0.67)
	Bal. Acc.	50.95	50.43	59.15 (±13.19)	41.90 (±1.32)	62.42 (±1.21)	50.48 (±1.26)	51.39 (±2.09)	85.05 (±2.27)	**92.46** (±1.61)
australian	Acc.	46.38	58.70	61.94 (±7.50)	41.28 (±1.26)	60.17 (±1.27)	62.84 (±1.79)	55.42 (±0.48)	78.26 (±0.68)	**81.97** (±1.09)
	AUROC	51.71	60.20	60.05 (±10.05)	43.47 (±1.03)	59.95 (±4.21)	66.02 (±1.13)	50.19 (±0.68)	85.24 (±0.23)	**87.84** (±0.56)
	AUPRC	55.52	58.67	57.94 (±7.87)	40.63 (±0.93)	56.55 (±3.79)	62.16 (±2.18)	45.57 (±3.33)	82.69 (±0.56)	**85.57** (±0.48)
	Bal. Acc.	48.21	57.11	61.76 (±7.62)	42.86 (±1.20)	58.70 (±1.80)	61.03 (±2.10)	50.28 (±0.66)	77.23 (±0.82)	**81.58** (±0.98)
wbc	Acc.	57.51	71.24	88.70 (±0.00)	25.73 (±0.83)	81.63 (±6.20)	85.38 (±1.52)	92.16 (±0.43)	93.02 (±0.67)	**95.22** (±0.33)
	AUROC	53.15	74.93	93.02 (±0.00)	37.69 (±0.44)	81.38 (±10.93)	90.46 (±1.50)	97.15 (±0.40)	98.11 (±0.10)	**98.39** (±0.04)
	AUPRC	64.01	69.06	91.05 (±0.00)	27.22 (±0.11)	80.36 (±10.54)	87.77 (±1.81)	95.25 (±0.48)	95.77 (±0.19)	**96.53** (±0.15)
	Bal. Acc.	46.84	66.46	87.84 (±0.00)	31.18 (±1.18)	80.52 (±3.75)	83.79 (±2.17)	91.13 (±0.65)	91.41 (±1.03)	**94.84** (±0.35)
diabetes_f	Acc.	63.54	63.80	63.10 (±6.32)	55.41 (±1.12)	62.68 (±1.24)	60.31 (±5.24)	64.32 (±0.13)	70.23 (±0.83)	**71.07** (±0.62)
	AUROC	49.90	59.49	61.37 (±9.84)	47.30 (±0.74)	47.43 (±1.82)	48.06 (±2.54)	50.13 (±0.27)	71.16 (±0.65)	**74.93** (±0.98)
	AUPRC	**65.60**	46.14	44.80 (±7.82)	32.72 (±0.64)	36.80 (±1.71)	34.83 (±1.66)	36.54 (±1.68)	57.57 (±1.07)	60.78 (±1.54)
	Bal. Acc.	50.10	49.17	58.57 (±5.57)	46.70 (±1.20)	51.81 (±0.51)	50.37 (±0.85)	50.21 (±0.30)	59.90 (±0.99)	**65.18** (±1.45)
german	Acc.	60.60	53.50	66.60 (±2.88)	69.97 (±0.00)	69.08 (±0.41)	64.56 (±4.00)	69.22 (±0.38)	70.12 (±0.58)	**72.72** (±0.54)
	AUROC	46.00	49.23	59.75 (±4.87)	48.28 (±0.40)	51.12 (±1.95)	47.22 (±2.67)	58.14 (±0.30)	63.04 (±0.54)	**71.70** (±0.60)
	AUPRC	71.81	69.01	75.60 (±3.09)	68.94 (±0.30)	71.30 (±1.97)	68.47 (±2.44)	74.31 (±0.14)	78.46 (±0.39)	**84.51** (±0.52)
	Bal. Acc.	54.05	51.17	54.77 (±2.19)	50.00 (±0.00)	49.82 (±0.26)	48.99 (±1.11)	51.96 (±0.35)	53.32 (±0.68)	**63.37** (±0.42)
svmguide3	Acc.	68.54	71.12	75.75 (±0.47)	75.96 (±0.15)	75.03 (±1.55)	73.72 (±2.56)	76.32 (±0.25)	76.11 (±0.43)	**78.65** (±0.55)
	AUROC	49.02	53.37	55.54 (±1.77)	46.60 (±0.41)	47.54 (±0.57)	48.80 (±1.13)	57.82 (±0.52)	65.82 (±1.04)	**68.60** (±1.34)
	AUPRC	76.93	75.74	78.75 (±1.38)	74.15 (±0.18)	73.65 (±0.67)	75.02 (±0.95)	79.15 (±0.27)	84.11 (±0.46)	**85.38** (±0.73)
	Bal. Acc.	51.02	53.64	51.27 (±0.57)	49.86 (±0.10)	50.28 (±0.42)	50.68 (±0.68)	50.73 (±0.33)	52.71 (±1.21)	**60.02** (±1.25)
LUdata	Acc.	58.81	68.33	80.54 (±0.00)	42.76 (±0.53)	51.87 (±0.80)	51.55 (±1.22)	73.08 (±0.33)	80.33 (±1.33)	**86.90** (±0.46)
	AUROC	41.15	56.03	89.33 (±0.00)	43.47 (±0.24)	53.47 (±3.46)	52.82 (±1.69)	77.23 (±0.16)	89.95 (±1.59)	**94.87** (±0.16)
	AUPRC	53.94	55.26	90.00 (±0.00)	43.85 (±0.16)	52.55 (±3.19)	51.40 (±2.23)	79.92 (±0.40)	89.45 (±1.68)	**94.90** (±0.13)
	Bal. Acc.	58.88	68.30	80.43 (±0.00)	42.19 (±0.54)	50.57 (±0.79)	50.88 (±0.91)	72.93 (±0.33)	80.21 (±1.34)	**86.87** (±0.43)
kr-vs-kp	Acc.	57.26	65.89	77.07 (±0.00)	42.46 (±0.22)	65.97 (±0.76)	64.45 (±0.60)	71.83 (±0.42)	76.46 (±1.28)	**83.37** (±0.22)
	AUROC	44.16	49.95	83.81 (±0.00)	42.79 (±0.09)	70.02 (±1.07)	69.71 (±1.12)	78.71 (±0.32)	84.79 (±0.97)	**91.64** (±0.18)
	AUPRC	46.45	52.66	82.49 (±0.00)	46.74 (±0.13)	70.46 (±0.90)	69.66 (±2.06)	79.82 (±0.33)	85.81 (±1.03)	**92.69** (±0.15)
	Bal. Acc.	55.81	65.74	76.92 (±0.00)	43.96 (±0.23)	65.74 (±0.74)	64.20 (±0.56)	71.71 (±0.42)	76.37 (±1.32)	**83.37** (±0.21)
spambase	Acc.	46.73	65.12	48.75 (±5.81)	60.18 (±0.10)	65.92 (±0.31)	66.47 (±0.81)	67.74 (±6.69)	85.72 (±0.39)	**89.90** (±0.16)
	AUROC	49.30	57.59	49.53 (±1.89)	44.77 (±0.41)	63.94 (±0.89)	66.73 (±1.68)	66.36 (±15.15)	91.72 (±0.35)	**95.02** (±0.11)
	AUPRC	48.22	50.04	55.03 (±11.65)	35.32 (±0.19)	57.11 (±0.92)	59.77 (±1.35)	54.75 (±16.89)	89.19 (±0.43)	**92.17** (±0.18)
	Bal. Acc.	50.65	59.80	49.96 (±1.86)	49.69 (±0.09)	58.84 (±0.32)	63.15 (±1.43)	61.75 (±10.73)	84.18 (±0.48)	**89.27** (±0.22)
spam	Acc.	74.96	94.51	91.44 (±0.00)	58.60 (±0.34)	86.04 (±2.74)	79.56 (±2.04)	95.38 (±0.26)	87.58 (±0.17)	**95.61** (±0.06)
	AUROC	24.93	35.96	96.12 (±0.00)	71.54 (±0.19)	93.46 (±1.91)	86.97 (±1.26)	98.39 (±0.08)	93.63 (±0.19)	**98.70** (±0.06)
	AUPRC	31.81	26.66	92.17 (±0.00)	64.91 (±0.10)	86.13 (±3.77)	72.63 (±2.12)	96.00 (±0.18)	87.93 (±0.20)	**96.94** (±0.10)
	Bal. Acc.	75.15	93.40	90.52 (±0.00)	63.52 (±0.32)	87.60 (±2.24)	81.21 (±1.83)	94.84 (±0.33)	85.15 (±0.20)	**95.16** (±0.09)
Synthetic datasets										
LED$_s$-A	Acc.	57.18	53.48	65.94 (±0.00)	39.36 (±0.15)	49.56 (±0.57)	51.53 (±1.62)	58.35 (±0.28)	71.63 (±0.53)	**73.68** (±0.59)
	AUROC	42.89	58.64	71.14 (±0.00)	45.63 (±0.07)	50.96 (±4.32)	51.64 (±2.39)	61.09 (±0.33)	78.62 (±0.58)	**81.71** (±0.48)
	AUPRC	60.21	57.42	69.38 (±0.00)	46.13 (±0.08)	50.52 (±3.35)	51.08 (±2.40)	59.26 (±0.23)	77.42 (±0.81)	**81.71** (±0.38)
	Bal. Acc.	57.11	53.65	65.94 (±0.00)	39.18 (±0.15)	49.85 (±0.60)	51.49 (±1.53)	58.36 (±0.28)	71.63 (±0.53)	**73.67** (±0.59)
LED$_s$-G	Acc.	56.80	54.14	66.56 (±0.00)	39.64 (±0.16)	49.63 (±0.45)	51.28 (±1.72)	58.64 (±0.47)	71.82 (±0.44)	**73.76** (±0.51)
	AUROC	43.27	57.99	71.81 (±0.00)	44.75 (±0.05)	50.62 (±4.37)	51.83 (±2.41)	61.30 (±0.67)	78.77 (±0.53)	**81.88** (±0.39)
	AUPRC	60.42	56.93	69.61 (±0.00)	45.57 (±0.04)	50.23 (±3.52)	50.86 (±1.92)	58.87 (±0.50)	77.68 (±0.49)	**81.79** (±0.44)
	Bal. Acc.	56.73	54.31	66.56 (±0.00)	39.45 (±0.16)	49.93 (±0.48)	51.22 (±1.69)	58.64 (±0.46)	71.81 (±0.43)	**73.75** (±0.52)
SEA$_s$-A	Acc.	52.70	60.36	71.14 (±0.07)	55.25 (±0.11)	58.67 (±1.86)	64.93 (±0.30)	70.95 (±0.48)	77.07 (±0.08)	**78.72** (±0.16)
	AUROC	47.98	71.75	77.46 (±0.01)	44.49 (±0.20)	63.65 (±8.88)	71.61 (±1.05)	76.16 (±0.40)	81.47 (±0.28)	**82.86** (±0.17)
	AUPRC	68.11	79.81	81.57 (±0.02)	55.38 (±0.13)	72.74 (±5.55)	78.65 (±0.68)	80.77 (±0.35)	83.77 (±0.27)	**84.82** (±0.12)
	Bal. Acc.	52.02	50.51	67.58 (±0.07)	46.44 (±0.06)	49.28 (±1.48)	58.22 (±0.36)	68.03 (±0.86)	74.84 (±0.17)	**76.76** (±0.12)
SEA$_s$-G	Acc.	52.74	60.30	71.11 (±0.04)	55.18 (±0.03)	58.77 (±1.71)	64.79 (±0.20)	70.53 (±0.88)	77.05 (±0.34)	**78.76** (±0.12)
	AUROC	48.05	71.60	77.21 (±0.02)	43.86 (±0.03)	63.29 (±8.97)	71.36 (±0.92)	75.34 (±1.28)	81.55 (±0.15)	**82.88** (±0.15)
	AUPRC	68.05	79.96	81.60 (±0.04)	54.72 (±0.12)	72.58 (±5.61)	78.66 (±0.71)	80.59 (±0.71)	84.48 (±0.13)	**85.33** (±0.09)
	Bal. Acc.	51.96	50.51	67.60 (±0.06)	46.48 (±0.02)	49.39 (±1.31)	58.23 (±0.37)	67.88 (±0.98)	74.97 (±0.49)	**76.86** (±0.07)
STAGGER$_s$-A	Acc.	72.60	67.02	79.14 (±0.00)	67.01 (±0.03)	57.77 (±3.81)	70.85 (±0.32)	82.15 (±0.41)	**89.46** (±0.25)	89.08 (±0.14)
	AUROC	31.75	59.08	86.61 (±0.00)	39.31 (±0.25)	43.16 (±0.09)	63.86 (±0.51)	89.95 (±0.31)	95.16 (±0.07)	**95.40** (±0.20)
	AUPRC	42.05	41.62	75.58 (±0.00)	26.89 (±0.15)	27.50 (±0.31)	52.07 (±0.85)	81.22 (±0.65)	92.67 (±0.13)	**93.17** (±0.22)
	Bal. Acc.	68.24	49.99	74.83 (±0.00)	49.99 (±0.03)	44.50 (±2.49)	58.15 (±0.71)	79.50 (±0.32)	85.85 (±0.14)	**85.97** (±0.13)
STAGGER$_s$-G	Acc.	71.74	67.02	78.70 (±0.00)	67.01 (±0.03)	57.80 (±3.84)	71.28 (±0.56)	80.61 (±0.62)	87.97 (±0.48)	**88.23** (±0.22)
	AUROC	33.01	58.69	85.25 (±0.00)	40.61 (±0.15)	42.56 (±0.73)	64.59 (±0.55)	88.33 (±0.36)	93.68 (±0.18)	**94.17** (±0.13)
	AUPRC	43.32	41.07	73.63 (±0.00)	27.64 (±0.10)	27.25 (±0.52)	52.82 (±1.04)	80.13 (±0.77)	90.73 (±0.36)	**91.47** (±0.11)
	Bal. Acc.	66.99	49.99	74.01 (±0.00)	49.99 (±0.03)	44.54 (±2.54)	59.01 (±1.06)	77.27 (±0.57)	83.70 (±0.37)	**85.02** (±0.02)

Table 8: Detailed results on small-scale streams ($n < 10{,}000$) at $p = 0.50$. Higher is better for all metrics. Bold and underlined values indicate the best and second-best results, respectively.

Dataset	Metric	Classical Online Learning				Online Deep Learning				
		OLIFL	OLVF	OLWDF	ORF^{3}V	HBP	Aux-Drop	MODL	HI^2	Our
Real-world datasets										
wdbc	Acc.	59.75	60.98	84.67 (±0.57)	44.93 (±1.07)	63.23 (±1.68)	62.14 (±1.64)	62.28 (±0.34)	84.43 (±0.80)	**88.40** (±1.77)
	AUROC	48.49	68.86	92.68 (±0.18)	37.63 (±1.13)	55.74 (±4.40)	49.30 (±2.05)	50.54 (±0.65)	92.68 (±0.40)	**94.70** (±0.60)
	AUPRC	62.14	61.23	86.70 (±0.42)	31.52 (±0.57)	47.19 (±3.01)	37.47 (±1.94)	41.68 (±4.64)	89.50 (±0.53)	**92.84** (±0.57)
	Bal. Acc.	51.55	53.48	83.30 (±0.75)	41.32 (±1.21)	59.55 (±1.41)	50.41 (±1.66)	50.21 (±0.56)	80.98 (±1.28)	**87.75** (±1.24)
australian	Acc.	47.97	57.39	72.46 (±0.31)	42.50 (±0.47)	56.70 (±1.07)	57.42 (±1.92)	55.45 (±0.43)	75.33 (±1.33)	**76.93** (±1.33)
	AUROC	49.29	58.29	79.85 (±0.19)	44.96 (±0.92)	58.04 (±2.15)	58.14 (±2.91)	51.35 (±0.35)	81.20 (±1.03)	**83.87** (±0.78)
	AUPRC	50.88	56.92	71.40 (±0.19)	42.26 (±0.81)	54.55 (±1.93)	54.56 (±2.72)	50.24 (±1.36)	76.80 (±1.26)	**80.88** (±0.81)
	Bal. Acc.	50.61	56.09	71.71 (±0.34)	44.11 (±0.65)	55.42 (±0.61)	55.80 (±1.82)	50.68 (±0.38)	74.03 (±1.53)	**76.38** (±1.50)
wbc	Acc.	57.22	56.94	87.84 (±0.00)	25.82 (±1.93)	78.66 (±6.85)	85.01 (±0.81)	91.42 (±0.40)	92.22 (±0.54)	**94.13** (±0.54)
	AUROC	49.32	76.70	93.80 (±0.00)	38.69 (±1.95)	77.28 (±10.61)	89.31 (±1.53)	95.76 (±0.26)	97.39 (±0.08)	**97.81** (±0.07)
	AUPRC	57.37	69.87	90.58 (±0.00)	27.93 (±0.82)	75.27 (±10.95)	85.96 (±1.74)	93.84 (±0.28)	95.06 (±0.14)	**95.81** (±0.32)
	Bal. Acc.	50.65	55.05	86.20 (±0.00)	31.82 (±2.22)	75.61 (±4.17)	82.27 (±1.69)	89.69 (±0.59)	90.37 (±0.79)	**93.38** (±0.49)
diabetes_f	Acc.	63.80	65.10	66.28 (±0.00)	58.90 (±1.35)	61.51 (±1.42)	59.45 (±6.72)	63.52 (±1.39)	67.81 (±0.21)	**68.85** (±0.63)
	AUROC	48.66	57.14	66.53 (±0.00)	50.24 (±0.81)	46.54 (±0.85)	50.44 (±0.65)	49.74 (±0.62)	64.80 (±0.61)	**70.07** (±0.50)
	AUPRC	**63.84**	42.17	51.28 (±0.00)	35.04 (±0.50)	33.89 (±0.98)	36.01 (±1.74)	36.24 (±0.97)	51.38 (±0.55)	54.99 (±1.65)
	Bal. Acc.	51.34	50.09	59.82 (±0.00)	49.16 (±1.16)	49.72 (±0.41)	51.20 (±0.90)	50.13 (±0.44)	55.53 (±0.37)	**60.98** (±1.23)
german	Acc.	60.10	51.30	70.28 (±0.28)	69.97 (±0.00)	69.04 (±0.34)	61.84 (±5.46)	67.70 (±0.23)	68.78 (±0.52)	**70.92** (±0.48)
	AUROC	46.26	49.37	63.08 (±0.46)	47.76 (±0.22)	49.97 (±1.70)	45.76 (±2.18)	52.51 (±0.81)	58.46 (±0.66)	**67.38** (±0.80)
	AUPRC	72.17	69.86	77.35 (±0.33)	68.25 (±0.25)	70.25 (±1.85)	67.55 (±1.79)	71.78 (±0.42)	75.27 (±0.36)	**81.74** (±0.71)
	Bal. Acc.	53.79	50.74	55.99 (±0.37)	50.00 (±0.00)	50.13 (±0.36)	49.16 (±1.29)	51.40 (±0.49)	50.71 (±0.40)	**59.65** (±0.60)
svmguide3	Acc.	71.52	73.05	76.32 (±0.12)	75.23 (±0.35)	75.08 (±1.53)	74.18 (±1.84)	76.09 (±0.15)	75.70 (±0.14)	**77.60** (±0.13)
	AUROC	47.77	50.28	60.58 (±0.54)	46.16 (±0.24)	47.88 (±0.48)	48.84 (±0.74)	55.72 (±0.46)	61.58 (±0.74)	**66.23** (±1.39)
	AUPRC	74.15	75.74	81.93 (±0.52)	74.69 (±0.27)	74.28 (±0.46)	75.09 (±0.77)	78.23 (±0.29)	81.48 (±0.67)	**84.06** (±0.67)
	Bal. Acc.	52.28	53.51	51.15 (±0.12)	49.71 (±0.24)	50.32 (±0.44)	50.61 (±0.45)	50.12 (±0.20)	51.08 (±0.66)	**57.66** (±0.90)
LUdata	Acc.	57.23	60.02	72.38 (±0.00)	43.18 (±0.74)	51.24 (±0.55)	50.38 (±1.47)	63.63 (±0.61)	69.26 (±1.23)	**76.82** (±0.77)
	AUROC	42.92	53.55	81.79 (±0.00)	45.47 (±0.45)	49.66 (±3.37)	49.62 (±1.10)	67.05 (±0.34)	77.82 (±1.94)	**86.80** (±0.30)
	AUPRC	57.66	53.15	82.66 (±0.00)	44.91 (±0.33)	49.14 (±3.64)	48.39 (±1.08)	69.54 (±0.23)	75.99 (±1.97)	**86.86** (±0.29)
	Bal. Acc.	57.10	59.93	72.16 (±0.00)	42.58 (±0.74)	49.88 (±0.49)	49.70 (±0.67)	63.37 (±0.59)	69.17 (±1.22)	**76.81** (±0.79)
kr-vs-kp	Acc.	54.29	57.57	69.99 (±0.00)	42.69 (±0.35)	58.60 (±0.73)	59.68 (±0.72)	63.27 (±0.54)	66.65 (±0.94)	**74.29** (±0.59)
	AUROC	47.18	50.87	75.67 (±0.00)	44.97 (±0.07)	60.65 (±1.31)	62.95 (±0.66)	68.19 (±0.20)	73.08 (±1.01)	**82.86** (±0.15)
	AUPRC	51.15	52.65	73.90 (±0.00)	47.93 (±0.08)	61.96 (±1.13)	63.53 (±0.59)	69.97 (±0.21)	74.19 (±1.27)	**84.35** (±0.18)
	Bal. Acc.	52.79	57.44	69.75 (±0.00)	44.18 (±0.38)	57.96 (±0.71)	59.32 (±0.73)	63.02 (±0.58)	66.44 (±0.96)	**74.23** (±0.54)
spambase	Acc.	48.69	66.38	69.39 (±0.19)	60.13 (±0.13)	66.00 (±0.41)	64.84 (±2.20)	61.62 (±2.74)	81.22 (±0.63)	**86.07** (±0.19)
	AUROC	45.48	58.07	74.46 (±0.09)	46.64 (±0.27)	61.39 (±0.62)	63.40 (±0.43)	53.92 (±8.92)	87.62 (±0.55)	**92.42** (±0.15)
	AUPRC	42.07	49.49	64.12 (±0.10)	37.05 (±0.16)	55.74 (±0.73)	57.16 (±1.05)	42.80 (±8.63)	84.45 (±0.68)	**88.49** (±0.33)
	Bal. Acc.	53.85	61.54	67.18 (±0.18)	49.73 (±0.11)	58.55 (±0.55)	60.80 (±0.76)	52.41 (±5.24)	79.02 (±0.70)	**85.13** (±0.22)
spam	Acc.	73.60	92.10	88.99 (±0.00)	57.99 (±0.19)	77.83 (±2.59)	74.76 (±1.58)	92.67 (±0.34)	88.07 (±0.17)	**94.40** (±0.22)
	AUROC	25.68	34.55	93.88 (±0.00)	71.37 (±0.12)	84.93 (±3.34)	80.16 (±1.19)	97.28 (±0.12)	93.90 (±0.08)	**98.16** (±0.05)
	AUPRC	31.57	25.15	89.63 (±0.00)	63.68 (±0.07)	71.05 (±5.43)	62.37 (±1.95)	94.03 (±0.27)	88.51 (±0.18)	**95.87** (±0.06)
	Bal. Acc.	74.33	90.55	87.91 (±0.00)	62.76 (±0.14)	78.11 (±3.69)	74.70 (±1.47)	91.23 (±0.43)	85.81 (±0.18)	**93.73** (±0.28)
Synthetic datasets										
LED$_s$-A	Acc.	54.60	54.08	61.58 (±0.00)	40.21 (±0.16)	50.27 (±0.98)	50.67 (±0.48)	54.68 (±0.22)	65.17 (±0.91)	**67.83** (±0.09)
	AUROC	45.46	51.75	66.43 (±0.00)	47.70 (±0.03)	50.72 (±2.30)	50.84 (±0.20)	56.11 (±0.15)	70.66 (±0.57)	**74.66** (±0.12)
	AUPRC	61.44	52.48	64.84 (±0.00)	47.28 (±0.05)	50.58 (±1.95)	50.46 (±0.51)	54.66 (±0.07)	69.47 (±0.72)	**74.56** (±0.08)
	Bal. Acc.	54.54	54.13	61.58 (±0.00)	40.05 (±0.16)	50.54 (±0.93)	50.61 (±0.51)	54.68 (±0.22)	65.16 (±0.90)	**67.81** (±0.09)
LED$_s$-G	Acc.	54.50	53.58	62.06 (±0.00)	41.13 (±0.21)	50.13 (±1.11)	50.37 (±0.50)	54.45 (±0.44)	66.30 (±0.44)	**68.21** (±0.40)
	AUROC	45.57	52.42	66.94 (±0.00)	46.89 (±0.02)	50.39 (±2.45)	50.45 (±0.88)	55.74 (±0.29)	71.29 (±0.42)	**75.03** (±0.20)
	AUPRC	61.58	52.31	65.34 (±0.00)	46.78 (±0.06)	50.35 (±2.04)	50.34 (±0.94)	54.51 (±0.10)	70.02 (±0.71)	**74.83** (±0.10)
	Bal. Acc.	54.44	53.61	62.05 (±0.00)	40.95 (±0.21)	50.40 (±1.08)	50.29 (±0.39)	54.45 (±0.44)	66.30 (±0.42)	**68.19** (±0.40)
SEA$_s$-A	Acc.	51.50	60.30	66.74 (±0.00)	51.10 (±0.50)	58.80 (±1.04)	61.30 (±0.35)	67.42 (±0.32)	70.99 (±0.48)	**72.98** (±0.23)
	AUROC	49.77	63.62	72.26 (±0.00)	44.03 (±0.34)	59.11 (±6.92)	66.15 (±0.83)	71.60 (±0.41)	75.47 (±0.45)	**77.44** (±0.20)
	AUPRC	68.82	73.83	78.27 (±0.00)	55.94 (±0.29)	69.91 (±4.23)	74.99 (±0.86)	77.35 (±0.20)	79.47 (±0.48)	**80.96** (±0.17)
	Bal. Acc.	50.24	50.03	62.67 (±0.00)	43.78 (±0.39)	49.21 (±0.84)	53.19 (±0.44)	63.68 (±0.40)	67.90 (±0.50)	**70.30** (±0.13)
SEA$_s$-G	Acc.	52.20	60.30	67.02 (±0.00)	50.82 (±0.10)	58.78 (±1.07)	61.12 (±0.08)	67.03 (±0.06)	70.59 (±0.40)	**72.68** (±0.21)
	AUROC	49.16	63.41	72.11 (±0.00)	43.99 (±0.27)	58.51 (±7.15)	65.92 (±0.70)	71.02 (±0.57)	75.54 (±0.37)	**77.47** (±0.25)
	AUPRC	68.27	73.92	78.27 (±0.00)	55.65 (±0.14)	69.55 (±4.15)	74.88 (±0.64)	77.37 (±0.34)	80.06 (±0.44)	**81.37** (±0.16)
	Bal. Acc.	50.85	50.03	62.80 (±0.00)	43.52 (±0.06)	49.21 (±0.85)	52.71 (±0.32)	63.36 (±0.35)	67.63 (±0.53)	**70.01** (±0.13)
STAGGER$_s$-A	Acc.	70.06	67.02	73.10 (±0.00)	67.03 (±0.01)	56.29 (±4.24)	68.56 (±0.29)	76.88 (±0.34)	81.01 (±0.53)	**81.08** (±0.13)
	AUROC	35.26	59.62	80.55 (±0.00)	41.67 (±0.51)	43.15 (±5.30)	60.69 (±0.55)	84.15 (±0.33)	88.26 (±0.19)	**88.87** (±0.29)
	AUPRC	45.42	40.36	66.79 (±0.00)	27.40 (±0.27)	28.76 (±4.33)	46.49 (±0.89)	71.59 (±0.78)	82.19 (±0.47)	**83.53** (±0.27)
	Bal. Acc.	64.73	49.99	70.17 (±0.00)	50.00 (±0.00)	43.65 (±2.50)	53.69 (±0.89)	72.97 (±0.41)	74.72 (±0.36)	**76.06** (±0.30)
STAGGER$_s$-G	Acc.	69.60	67.02	72.30 (±0.00)	67.03 (±0.01)	56.35 (±4.23)	68.52 (±0.17)	75.90 (±0.25)	**80.82** (±0.50)	80.77 (±0.05)
	AUROC	36.25	59.43	79.30 (±0.00)	42.93 (±0.54)	43.46 (±6.54)	60.36 (±0.32)	82.82 (±0.30)	86.76 (±0.20)	**87.69** (±0.22)
	AUPRC	46.65	40.71	64.53 (±0.00)	28.15 (±0.34)	29.55 (±5.92)	46.37 (±0.45)	71.61 (±0.54)	80.81 (±0.23)	**82.23** (±0.24)
	Bal. Acc.	63.74	49.99	66.99 (±0.00)	50.00 (±0.00)	43.70 (±2.52)	53.67 (±0.56)	70.80 (±0.22)	73.69 (±0.50)	**75.97** (±0.31)

Table 9: Detailed results on small-scale streams ($n < 10{,}000$) at $p = 0.25$. Higher is better for all metrics. Bold and underlined values indicate the best and second-best results, respectively.

Dataset	Metric	Classical Online Learning				Online Deep Learning				
		OLIFL	OLVF	OLWDF	ORF^{3}V	HBP	Aux-Drop	MODL	HI^2	Our
Real-world datasets										
wdbc	Acc.	58.35	56.41	81.34 (±0.45)	44.26 (±1.89)	57.72 (±2.33)	61.05 (±1.16)	62.00 (±0.23)	80.25 (±0.16)	**82.88** (±0.95)
	AUROC	51.46	64.74	89.20 (±0.23)	34.78 (±1.16)	50.46 (±4.03)	45.41 (±1.99)	50.12 (±0.27)	88.13 (±0.52)	**89.29** (±0.65)
	AUPRC	65.76	60.37	84.01 (±0.58)	28.65 (±0.47)	39.95 (±2.96)	34.73 (±0.79)	38.97 (±2.49)	80.91 (±1.38)	**84.30** (±0.62)
	Bal. Acc.	48.61	49.17	79.30 (±0.60)	41.32 (±1.76)	52.68 (±1.83)	49.40 (±0.32)	49.83 (±0.31)	76.33 (±0.49)	**80.71** (±0.97)
australian	Acc.	51.30	54.49	**66.75** (±0.43)	43.83 (±0.79)	53.68 (±3.24)	52.38 (±3.39)	54.70 (±0.64)	63.39 (±1.19)	66.52 (±1.74)
	AUROC	46.06	60.30	71.43 (±0.08)	49.23 (±0.23)	55.43 (±1.50)	50.14 (±3.40)	50.52 (±0.39)	67.50 (±0.64)	**72.03** (±1.04)
	AUPRC	49.11	55.66	65.28 (±0.05)	44.71 (±0.23)	51.69 (±0.83)	46.99 (±2.38)	47.12 (±1.19)	62.49 (±0.88)	**68.58** (±0.94)
	Bal. Acc.	53.42	53.09	65.23 (±0.40)	45.11 (±0.86)	52.91 (±1.82)	50.76 (±1.99)	50.04 (±0.74)	60.82 (±1.34)	**65.55** (±1.54)
wbc	Acc.	41.77	48.07	83.55 (±0.00)	24.47 (±0.81)	75.28 (±4.12)	82.09 (±0.92)	87.75 (±0.28)	87.24 (±0.67)	**88.70** (±0.56)
	AUROC	52.10	80.25	87.92 (±0.00)	41.21 (±0.91)	71.66 (±8.53)	85.02 (±0.58)	91.10 (±0.31)	93.45 (±0.08)	**93.91** (±0.14)
	AUPRC	46.67	74.48	84.22 (±0.00)	28.99 (±0.65)	67.29 (±9.82)	80.57 (±1.02)	88.23 (±0.41)	88.91 (±0.12)	**89.82** (±0.21)
	Bal. Acc.	47.90	52.02	80.56 (±0.00)	29.38 (±1.13)	68.49 (±2.39)	77.07 (±1.32)	84.50 (±0.22)	83.85 (±0.98)	**86.46** (±0.47)
diabetes_f	Acc.	58.33	64.45	66.93 (±0.00)	49.52 (±0.56)	60.23 (±1.94)	55.86 (±8.43)	63.31 (±1.08)	66.46 (±0.59)	**67.01** (±1.37)
	AUROC	50.16	55.49	61.91 (±0.00)	50.62 (±0.87)	46.29 (±1.01)	49.15 (±1.22)	50.57 (±0.95)	61.92 (±0.68)	**64.14** (±1.10)
	AUPRC	**60.45**	39.27	49.34 (±0.00)	35.04 (±0.49)	34.15 (±0.86)	35.56 (±1.10)	37.04 (±2.19)	48.66 (±0.61)	49.37 (±2.40)
	Bal. Acc.	49.82	49.50	**58.15** (±0.00)	45.74 (±0.31)	48.68 (±0.87)	49.51 (±0.70)	50.49 (±0.61)	53.12 (±0.85)	55.96 (±1.65)
german	Acc.	56.60	44.40	67.20 (±0.00)	**69.77** (±0.19)	69.18 (±0.54)	57.84 (±9.56)	68.98 (±0.43)	68.60 (±0.66)	69.28 (±0.60)
	AUROC	47.00	49.02	55.90 (±0.00)	47.48 (±0.19)	47.38 (±1.09)	46.48 (±2.24)	53.54 (±0.41)	52.02 (±0.21)	**59.86** (±0.83)
	AUPRC	74.20	69.08	73.96 (±0.00)	68.28 (±0.23)	68.28 (±1.07)	67.56 (±1.57)	71.48 (±0.27)	70.95 (±0.23)	**77.58** (±0.26)
	Bal. Acc.	**53.19**	48.57	50.00 (±0.00)	50.05 (±0.16)	49.74 (±0.31)	49.09 (±2.26)	50.40 (±0.49)	49.67 (±0.29)	52.51 (±0.39)
svmguide3	Acc.	52.05	71.76	76.03 (±0.06)	72.74 (±0.24)	75.00 (±1.64)	74.27 (±1.50)	76.09 (±0.07)	75.43 (±0.41)	**76.75** (±0.81)
	AUROC	48.51	53.45	56.33 (±0.25)	48.06 (±0.14)	48.11 (±0.29)	48.54 (±0.65)	52.48 (±0.20)	54.86 (±0.94)	**58.23** (±1.50)
	AUPRC	**81.30**	77.23	79.32 (±0.21)	75.08 (±0.12)	74.89 (±0.27)	74.97 (±0.70)	77.25 (±0.17)	78.42 (±0.49)	79.48 (±0.75)
	Bal. Acc.	51.69	52.09	49.89 (±0.04)	49.54 (±0.43)	50.26 (±0.37)	50.21 (±0.40)	50.01 (±0.14)	50.66 (±0.41)	**53.37** (±0.93)
LUdata	Acc.	52.13	52.13	62.81 (±0.15)	43.44 (±0.66)	51.15 (±0.49)	50.06 (±1.61)	56.23 (±0.67)	59.11 (±1.23)	**66.35** (±0.62)
	AUROC	47.54	54.37	68.86 (±0.05)	48.38 (±0.29)	47.54 (±1.51)	47.54 (±0.91)	57.63 (±0.26)	62.74 (±2.28)	**74.04** (±0.35)
	AUPRC	57.91	53.29	69.66 (±0.04)	46.86 (±0.26)	47.37 (±1.89)	46.63 (±0.82)	58.55 (±0.51)	60.36 (±2.62)	**74.17** (±0.16)
	Bal. Acc.	52.30	52.06	62.45 (±0.15)	43.11 (±0.64)	49.78 (±0.45)	49.36 (±0.80)	55.87 (±0.63)	58.89 (±1.22)	**66.30** (±0.67)
kr-vs-kp	Acc.	52.82	52.66	60.70 (±0.00)	43.58 (±0.19)	52.50 (±0.41)	54.12 (±0.35)	56.16 (±0.34)	57.25 (±0.31)	**63.94** (±0.49)
	AUROC	48.17	47.82	63.76 (±0.00)	47.07 (±0.14)	52.37 (±1.05)	54.47 (±0.18)	57.83 (±0.21)	58.71 (±0.40)	**69.99** (±0.24)
	AUPRC	56.64	50.51	62.72 (±0.00)	49.60 (±0.06)	54.22 (±1.24)	55.91 (±0.37)	58.88 (±0.29)	58.73 (±0.58)	**71.48** (±0.39)
	Bal. Acc.	51.81	52.47	60.39 (±0.00)	44.16 (±0.18)	51.22 (±0.42)	53.12 (±0.38)	55.80 (±0.24)	56.83 (±0.30)	**63.63** (±0.50)
spambase	Acc.	50.10	67.05	63.03 (±0.24)	58.67 (±0.12)	63.08 (±0.35)	62.35 (±1.57)	62.96 (±2.40)	74.56 (±0.73)	**80.52** (±0.34)
	AUROC	42.20	58.47	65.68 (±0.19)	46.70 (±0.26)	55.84 (±0.29)	57.17 (±0.95)	56.33 (±6.75)	79.45 (±0.93)	**86.97** (±0.06)
	AUPRC	40.96	48.46	53.93 (±0.31)	37.04 (±0.25)	49.19 (±0.36)	50.32 (±1.12)	46.20 (±7.37)	73.06 (±1.15)	**81.18** (±0.20)
	Bal. Acc.	54.87	64.34	59.88 (±0.29)	48.94 (±0.09)	54.33 (±0.23)	55.53 (±0.77)	55.40 (±4.91)	71.46 (±0.85)	**79.03** (±0.27)
spam	Acc.	72.69	86.56	85.80 (±0.00)	53.82 (±0.17)	71.56 (±0.12)	70.57 (±0.46)	87.69 (±0.47)	86.02 (±0.22)	**90.99** (±0.28)
	AUROC	25.41	35.68	90.21 (±0.00)	71.65 (±0.09)	70.24 (±0.33)	68.99 (±0.22)	93.67 (±0.37)	91.64 (±0.18)	**96.04** (±0.09)
	AUPRC	30.42	25.32	83.83 (±0.00)	62.22 (±0.04)	53.35 (±0.32)	51.21 (±0.38)	88.63 (±0.53)	85.62 (±0.31)	**92.51** (±0.21)
	Bal. Acc.	74.22	84.65	82.69 (±0.00)	59.26 (±0.20)	68.74 (±0.40)	69.15 (±0.53)	84.31 (±0.82)	83.08 (±0.37)	**89.38** (±0.45)
Synthetic datasets										
LED$_s$-A	Acc.	52.20	52.32	57.88 (±0.00)	41.31 (±0.25)	49.93 (±0.66)	49.82 (±0.51)	52.21 (±0.47)	57.47 (±1.02)	**62.41** (±0.15)
	AUROC	47.85	51.97	61.44 (±0.00)	47.86 (±0.05)	50.67 (±0.39)	49.55 (±0.19)	52.52 (±0.23)	60.80 (±0.80)	**67.25** (±0.28)
	AUPRC	62.62	51.93	60.17 (±0.00)	47.70 (±0.04)	50.36 (±0.53)	49.57 (±0.26)	51.85 (±0.32)	58.69 (±1.13)	**66.39** (±0.13)
	Bal. Acc.	52.15	52.34	57.86 (±0.00)	41.19 (±0.25)	50.16 (±0.53)	49.75 (±0.36)	52.20 (±0.46)	57.47 (±1.02)	**62.40** (±0.15)
LED$_s$-G	Acc.	52.48	52.50	57.52 (±0.00)	42.28 (±0.12)	49.37 (±0.31)	49.75 (±0.59)	52.27 (±0.47)	57.57 (±1.40)	**62.55** (±0.27)
	AUROC	47.58	51.23	61.69 (±0.00)	47.14 (±0.02)	50.38 (±0.27)	49.48 (±0.04)	52.30 (±0.38)	60.86 (±1.33)	**67.20** (±0.24)
	AUPRC	62.60	51.27	60.43 (±0.00)	47.35 (±0.03)	50.27 (±0.67)	49.37 (±0.21)	51.59 (±0.39)	58.87 (±1.59)	**66.26** (±0.16)
	Bal. Acc.	52.43	52.50	57.50 (±0.00)	42.12 (±0.12)	49.60 (±0.44)	49.66 (±0.42)	52.27 (±0.45)	57.57 (±1.40)	**62.54** (±0.27)
SEA$_s$-A	Acc.	52.22	60.26	65.58 (±0.00)	50.19 (±0.72)	59.33 (±0.34)	60.94 (±0.24)	65.97 (±0.25)	66.73 (±0.54)	**68.62** (±0.06)
	AUROC	49.39	61.46	69.28 (±0.00)	44.69 (±0.49)	57.79 (±6.13)	64.69 (±0.74)	68.57 (±0.28)	69.98 (±0.63)	**72.16** (±0.13)
	AUPRC	68.20	71.54	76.24 (±0.00)	57.20 (±0.37)	69.49 (±3.38)	73.56 (±0.47)	75.31 (±0.30)	76.05 (±0.59)	**78.40** (±0.08)
	Bal. Acc.	50.63	49.98	60.36 (±0.00)	43.49 (±0.51)	49.56 (±0.47)	51.55 (±0.30)	61.93 (±0.34)	62.73 (±0.70)	**64.86** (±0.10)
SEA$_s$-G	Acc.	52.60	60.28	65.42 (±0.00)	49.65 (±0.65)	59.32 (±0.34)	60.53 (±0.26)	66.42 (±0.09)	66.79 (±0.71)	**68.66** (±0.14)
	AUROC	49.41	61.13	68.55 (±0.00)	44.39 (±0.24)	57.23 (±6.09)	64.26 (±0.67)	68.35 (±0.17)	69.88 (±0.63)	**72.12** (±0.17)
	AUPRC	67.82	71.20	75.54 (±0.00)	56.43 (±0.13)	69.01 (±3.27)	73.32 (±0.27)	75.21 (±0.03)	76.23 (±0.60)	**78.33** (±0.11)
	Bal. Acc.	50.61	50.00	59.97 (±0.00)	43.17 (±0.48)	49.55 (±0.47)	50.97 (±0.33)	62.53 (±0.31)	62.71 (±0.95)	**64.80** (±0.17)
STAGGER$_s$-A	Acc.	69.62	67.00	74.56 (±0.00)	66.84 (±0.08)	55.81 (±5.45)	67.79 (±0.29)	75.13 (±0.40)	**77.81** (±0.35)	76.95 (±0.37)
	AUROC	35.75	60.21	81.09 (±0.00)	41.93 (±0.15)	41.81 (±8.74)	60.12 (±0.85)	82.64 (±0.11)	83.94 (±0.14)	**84.98** (±0.10)
	AUPRC	45.79	40.55	65.28 (±0.00)	27.23 (±0.03)	30.75 (±9.67)	44.29 (±0.72)	69.65 (±0.28)	76.19 (±0.38)	**77.61** (±0.14)
	Bal. Acc.	64.23	49.97	**71.91** (±0.00)	49.88 (±0.08)	43.24 (±3.20)	52.27 (±0.07)	71.52 (±0.37)	70.18 (±0.25)	71.87 (±0.10)
STAGGER$_s$-G	Acc.	69.14	67.00	72.92 (±0.00)	66.91 (±0.05)	55.79 (±5.46)	67.49 (±0.30)	74.73 (±0.43)	**76.97** (±0.31)	76.63 (±0.39)
	AUROC	36.98	60.87	78.67 (±0.00)	43.95 (±0.28)	41.70 (±9.85)	58.03 (±0.50)	81.22 (±0.28)	82.29 (±0.08)	**83.47** (±0.11)
	AUPRC	47.42	40.42	61.02 (±0.00)	28.28 (±0.09)	30.94 (±10.39)	42.13 (±0.23)	68.94 (±0.59)	74.04 (±0.68)	**75.91** (±0.18)
	Bal. Acc.	63.01	49.97	68.05 (±0.00)	49.91 (±0.05)	43.22 (±3.21)	51.43 (±0.35)	70.13 (±0.76)	69.19 (±0.26)	**71.09** (±0.22)

Table 10: Detailed results on large-scale streams ($n \geq 10{,}000$) at $p = 0.75$. Higher is better for all metrics. Bold and underlined values indicate the best and second-best results, respectively.

Dataset	Metric	Classical Online Learning				Online Deep Learning				
		OLIFL	OLVF	OLWDF	ORF^{3}V	HBP	Aux-Drop	MODL	HI^2	Our
Real-world datasets										
phishing	Acc.	64.59	89.57	89.39 (±0.00)	54.98 (±0.02)	84.42 (±0.44)	79.31 (±0.98)	90.12 (±0.10)	89.01 (±0.27)	**92.31** (±0.12)
	AUROC	39.88	52.16	95.94 (±0.00)	42.19 (±0.09)	93.80 (±0.27)	90.14 (±0.66)	96.48 (±0.02)	95.79 (±0.14)	**97.91** (±0.03)
	AUPRC	29.78	54.78	96.56 (±0.00)	50.98 (±0.07)	94.98 (±0.20)	91.65 (±0.50)	97.12 (±0.02)	96.40 (±0.14)	**98.28** (±0.04)
	Bal. Acc.	60.11	89.14	89.07 (±0.00)	49.39 (±0.02)	83.36 (±0.57)	77.33 (±1.20)	89.87 (±0.11)	88.77 (±0.28)	**92.18** (±0.13)
weather	Acc.	66.50	64.36	53.91 (±3.02)	68.62 (±0.00)	68.60 (±0.05)	68.49 (±0.16)	68.56 (±0.06)	**75.22** (±0.23)	74.55 (±0.14)
	AUROC	42.17	45.44	47.57 (±0.27)	42.49 (±0.22)	54.68 (±0.70)	48.58 (±0.49)	50.04 (±0.06)	**79.04** (±0.29)	78.81 (±0.11)
	AUPRC	51.62	28.60	38.28 (±5.41)	26.36 (±0.12)	34.96 (±0.80)	30.26 (±0.71)	33.49 (±2.27)	**62.99** (±0.42)	61.86 (±0.32)
	Bal. Acc.	57.83	57.92	49.57 (±0.77)	50.00 (±0.00)	50.11 (±0.08)	50.02 (±0.04)	50.03 (±0.03)	67.48 (±0.47)	**68.73** (±0.14)
magic04	Acc.	62.67	62.31	70.22 (±0.30)	34.82 (±0.03)	73.72 (±0.12)	70.88 (±0.22)	73.13 (±0.23)	**78.80** (±0.33)	78.60 (±0.16)
	AUROC	39.24	46.89	70.86 (±0.47)	44.26 (±0.13)	72.73 (±0.35)	68.42 (±0.45)	72.11 (±0.22)	83.99 (±0.38)	**84.24** (±0.08)
	AUPRC	64.53	61.40	79.09 (±0.40)	59.68 (±0.14)	78.78 (±0.21)	76.11 (±0.55)	77.47 (±0.21)	89.46 (±0.27)	**89.64** (±0.10)
	Bal. Acc.	60.75	56.19	62.58 (±0.30)	49.32 (±0.04)	66.36 (±0.08)	63.29 (±0.83)	65.21 (±0.38)	74.03 (±0.28)	**74.48** (±0.17)
a8a	Acc.	77.67	79.66	82.63 (±0.00)	75.88 (±0.00)	81.71 (±0.22)	79.94 (±0.34)	80.27 (±0.15)	81.79 (±0.12)	**82.93** (±0.03)
	AUROC	43.44	72.43	87.03 (±0.00)	46.92 (±0.03)	84.35 (±0.67)	79.66 (±1.14)	84.16 (±0.25)	85.55 (±0.10)	**87.96** (±0.03)
	AUPRC	55.82	90.00	95.47 (±0.00)	74.06 (±0.02)	94.46 (±0.26)	92.35 (±0.43)	94.17 (±0.12)	94.80 (±0.05)	**95.75** (±0.02)
	Bal. Acc.	56.56	70.63	71.86 (±0.00)	49.99 (±0.00)	67.64 (±0.54)	62.88 (±0.94)	63.31 (±0.50)	70.31 (±0.19)	**76.65** (±0.08)
elec	Acc.	61.31	57.98	65.26 (±0.18)	44.31 (±0.18)	66.65 (±0.55)	60.98 (±0.64)	72.54 (±0.23)	**78.62** (±0.54)	77.72 (±0.45)
	AUROC	41.18	45.68	69.14 (±0.33)	46.31 (±0.18)	70.51 (±0.77)	62.60 (±0.79)	78.74 (±0.26)	86.28 (±0.62)	**86.40** (±0.41)
	AUPRC	55.87	38.53	60.61 (±0.07)	40.09 (±0.14)	64.62 (±0.91)	54.92 (±0.99)	74.39 (±0.36)	82.44 (±0.76)	**83.00** (±0.51)
	Bal. Acc.	58.82	54.49	62.43 (±0.18)	46.25 (±0.16)	63.26 (±0.72)	55.90 (±1.06)	70.85 (±0.28)	**77.84** (±0.57)	77.15 (±0.43)
Synthetic datasets										
LED$_m$-A	Acc.	59.73	51.56	67.80 (±0.00)	35.17 (±0.07)	50.88 (±0.30)	60.98 (±1.20)	64.27 (±0.04)	76.64 (±0.06)	**77.33** (±0.12)
	AUROC	40.30	59.75	74.06 (±0.00)	46.60 (±0.01)	52.41 (±2.55)	65.62 (±1.56)	69.73 (±0.11)	84.38 (±0.05)	**85.53** (±0.06)
	AUPRC	55.34	61.23	73.88 (±0.00)	48.09 (±0.01)	52.44 (±2.33)	66.20 (±2.66)	69.06 (±0.14)	84.05 (±0.08)	**85.69** (±0.05)
	Bal. Acc.	59.71	50.86	67.78 (±0.00)	35.28 (±0.07)	50.14 (±0.32)	60.92 (±1.26)	64.27 (±0.04)	76.63 (±0.06)	**77.32** (±0.12)
LED$_m$-G	Acc.	59.60	51.55	67.74 (±0.00)	35.22 (±0.12)	50.90 (±0.39)	60.51 (±1.52)	64.03 (±0.14)	76.07 (±0.06)	**77.09** (±0.16)
	AUROC	40.42	59.58	74.12 (±0.00)	46.68 (±0.02)	52.32 (±2.46)	64.87 (±2.03)	69.17 (±0.20)	83.78 (±0.07)	**85.19** (±0.10)
	AUPRC	55.46	60.94	74.19 (±0.00)	48.08 (±0.01)	52.38 (±2.31)	65.49 (±2.60)	68.60 (±0.25)	83.45 (±0.12)	**85.25** (±0.09)
	Bal. Acc.	59.58	50.88	67.71 (±0.00)	35.33 (±0.12)	50.20 (±0.42)	60.45 (±1.58)	64.03 (±0.15)	76.05 (±0.06)	**77.09** (±0.15)
SEA$_m$-A	Acc.	53.41	62.68	72.19 (±0.04)	56.39 (±0.15)	62.02 (±1.67)	72.03 (±0.50)	74.60 (±0.14)	79.42 (±0.09)	**80.07** (±0.07)
	AUROC	47.42	74.42	78.30 (±0.07)	46.22 (±0.09)	67.57 (±5.67)	77.95 (±0.59)	80.17 (±0.06)	84.70 (±0.06)	**85.01** (±0.02)
	AUPRC	67.70	81.60	83.10 (±0.04)	57.00 (±0.09)	76.33 (±3.01)	82.72 (±0.31)	83.78 (±0.15)	86.31 (±0.10)	**86.73** (±0.01)
	Bal. Acc.	52.58	54.60	68.98 (±0.02)	47.14 (±0.13)	52.79 (±2.80)	69.64 (±0.26)	72.41 (±0.25)	77.40 (±0.12)	**78.18** (±0.01)
SEA$_m$-G	Acc.	53.24	62.29	71.51 (±0.04)	56.31 (±0.12)	61.82 (±1.45)	71.37 (±0.57)	73.42 (±0.33)	78.60 (±0.17)	**79.33** (±0.02)
	AUROC	47.49	74.38	77.73 (±0.00)	45.36 (±0.06)	67.33 (±5.78)	77.08 (±0.59)	78.67 (±0.56)	83.81 (±0.09)	**84.21** (±0.01)
	AUPRC	67.86	81.32	82.57 (±0.00)	56.41 (±0.05)	76.08 (±3.03)	81.74 (±0.17)	82.66 (±0.34)	85.70 (±0.09)	**86.12** (±0.03)
	Bal. Acc.	52.51	54.09	68.19 (±0.05)	47.04 (±0.11)	52.54 (±2.51)	68.83 (±0.31)	71.02 (±0.22)	76.48 (±0.22)	**77.43** (±0.02)
STAGGER$_m$-A	Acc.	70.81	67.78	76.72 (±0.00)	60.00 (±0.01)	70.97 (±1.97)	79.54 (±0.82)	89.19 (±0.14)	**91.45** (±0.06)	90.65 (±0.13)
	AUROC	30.39	33.07	84.68 (±0.00)	43.09 (±0.18)	73.86 (±2.10)	83.65 (±0.60)	96.91 (±0.06)	97.39 (±0.03)	**97.68** (±0.02)
	AUPRC	43.14	29.99	78.20 (±0.00)	35.22 (±0.12)	64.29 (±2.99)	75.82 (±1.78)	95.86 (±0.08)	96.65 (±0.04)	**96.94** (±0.03)
	Bal. Acc.	69.61	62.40	75.79 (±0.00)	50.11 (±0.02)	70.52 (±3.03)	78.27 (±1.07)	88.57 (±0.15)	**90.29** (±0.05)	90.04 (±0.11)
STAGGER$_m$-G	Acc.	69.04	68.07	76.09 (±0.00)	59.93 (±0.01)	69.87 (±1.30)	79.44 (±1.53)	86.71 (±0.04)	**88.70** (±0.04)	88.08 (±0.07)
	AUROC	32.41	34.17	84.03 (±0.00)	45.70 (±0.28)	71.04 (±2.40)	84.05 (±1.35)	94.55 (±0.03)	95.64 (±0.04)	**95.95** (±0.02)
	AUPRC	45.11	30.48	78.23 (±0.00)	37.14 (±0.26)	62.31 (±3.26)	80.45 (±1.66)	92.86 (±0.04)	94.31 (±0.03)	**94.68** (±0.02)
	Bal. Acc.	67.59	62.91	74.81 (±0.00)	50.12 (±0.02)	68.52 (±2.63)	77.41 (±1.67)	85.76 (±0.05)	87.49 (±0.08)	**87.55** (±0.03)

Table 11: Detailed results on large-scale streams ($n \geq 10{,}000$) at $p = 0.50$. Higher is better for all metrics. Bold and underlined values indicate the best and second-best results, respectively.

Dataset	Metric	Classical Online Learning				Online Deep Learning				
		OLIFL	OLVF	OLWDF	ORF^{3}V	HBP	Aux-Drop	MODL	HI^2	Our
Real-world datasets										
phishing	Acc.	57.30	83.88	<u>85.82</u> (±0.00)	53.36 (±0.06)	76.85 (±0.71)	70.31 (±1.44)	84.66 (±0.25)	84.32 (±0.47)	**88.76** (±0.22)
	AUROC	48.11	54.88	<u>93.00</u> (±0.00)	43.37 (±0.09)	87.25 (±0.71)	81.47 (±1.37)	92.50 (±0.10)	92.04 (±0.21)	**95.60** (±0.05)
	AUPRC	34.88	57.47	<u>93.93</u> (±0.00)	51.16 (±0.07)	89.63 (±0.53)	84.82 (±1.07)	93.75 (±0.07)	93.03 (±0.19)	**96.31** (±0.06)
	Bal. Acc.	51.88	83.26	<u>85.44</u> (±0.00)	47.98 (±0.06)	75.13 (±0.86)	67.30 (±1.78)	84.27 (±0.26)	83.91 (±0.52)	**88.51** (±0.24)
weather	Acc.	66.10	62.76	69.51 (±0.28)	68.55 (±0.01)	68.58 (±0.08)	68.56 (±0.07)	68.59 (±0.02)	<u>71.44</u> (±0.34)	**71.80** (±0.15)
	AUROC	45.70	46.84	66.25 (±0.55)	46.22 (±0.10)	55.28 (±0.91)	47.07 (±0.12)	50.00 (±0.01)	<u>72.34</u> (±0.70)	**74.10** (±0.11)
	AUPRC	**56.82**	29.09	46.45 (±0.71)	28.21 (±0.05)	35.36 (±0.85)	29.09 (±0.25)	32.00 (±2.29)	53.18 (±0.81)	<u>54.89</u> (±0.19)
	Bal. Acc.	54.30	56.19	56.53 (±0.16)	49.98 (±0.01)	50.10 (±0.07)	50.00 (±0.02)	50.00 (±0.01)	<u>60.36</u> (±0.85)	**63.91** (±0.18)
magic04	Acc.	58.83	61.09	68.12 (±0.05)	35.04 (±0.09)	71.16 (±0.07)	68.92 (±0.24)	71.18 (±0.20)	<u>73.94</u> (±0.47)	**74.57** (±0.08)
	AUROC	42.71	45.39	65.90 (±0.05)	47.34 (±0.06)	66.53 (±0.25)	62.36 (±0.74)	67.25 (±0.26)	<u>76.35</u> (±0.65)	**77.91** (±0.13)
	AUPRC	67.54	60.30	76.11 (±0.04)	62.27 (±0.08)	74.06 (±0.21)	71.20 (±0.57)	74.24 (±0.28)	<u>84.16</u> (±0.51)	**85.38** (±0.11)
	Bal. Acc.	57.28	54.63	58.70 (±0.06)	48.57 (±0.08)	61.90 (±0.08)	59.27 (±0.45)	62.43 (±0.28)	<u>66.99</u> (±0.62)	**68.23** (±0.09)
a8a	Acc.	76.47	76.93	<u>80.79</u> (±0.00)	75.91 (±0.00)	79.42 (±0.33)	77.35 (±0.55)	77.87 (±0.19)	79.91 (±0.11)	**81.15** (±0.09)
	AUROC	45.36	69.63	<u>84.18</u> (±0.00)	47.29 (±0.06)	79.39 (±1.25)	72.34 (±2.56)	80.51 (±0.23)	81.69 (±0.10)	**85.45** (±0.03)
	AUPRC	60.82	88.60	<u>94.32</u> (±0.00)	74.27 (±0.03)	92.39 (±0.51)	89.27 (±0.93)	92.47 (±0.12)	93.09 (±0.08)	**94.67** (±0.02)
	Bal. Acc.	54.63	66.46	<u>67.65</u> (±0.00)	50.01 (±0.00)	61.56 (±0.86)	55.34 (±2.01)	56.65 (±0.64)	65.63 (±0.12)	**73.44** (±0.07)
elec	Acc.	57.00	54.90	65.23 (±0.02)	45.34 (±0.16)	62.23 (±0.44)	58.79 (±0.27)	67.71 (±0.11)	<u>73.26</u> (±0.66)	**73.86** (±0.36)
	AUROC	45.69	49.61	70.10 (±0.02)	47.33 (±0.07)	64.14 (±0.61)	57.55 (±0.73)	72.24 (±0.17)	<u>80.65</u> (±0.99)	**82.09** (±0.30)
	AUPRC	59.19	41.15	60.93 (±0.05)	40.65 (±0.07)	56.86 (±0.70)	49.87 (±0.80)	66.62 (±0.18)	<u>75.45</u> (±1.46)	**78.08** (±0.35)
	Bal. Acc.	54.31	51.02	63.11 (±0.03)	46.22 (±0.17)	57.63 (±0.64)	52.30 (±0.58)	65.27 (±0.15)	<u>72.12</u> (±0.64)	**72.87** (±0.35)
Synthetic datasets										
LED$_m$-A	Acc.	55.67	51.33	64.01 (±0.00)	37.88 (±0.11)	51.09 (±0.60)	55.53 (±1.18)	58.27 (±0.29)	<u>68.87</u> (±0.16)	**70.29** (±0.09)
	AUROC	44.36	54.66	69.45 (±0.00)	47.67 (±0.01)	51.44 (±1.54)	58.04 (±1.58)	61.68 (±0.38)	<u>75.66</u> (±0.12)	**77.68** (±0.08)
	AUPRC	58.40	56.10	69.16 (±0.00)	48.68 (±0.01)	52.04 (±1.56)	58.83 (±1.70)	61.62 (±0.18)	<u>75.20</u> (±0.14)	**77.71** (±0.10)
	Bal. Acc.	55.65	50.63	63.97 (±0.00)	37.98 (±0.11)	50.42 (±0.73)	55.42 (±1.21)	58.27 (±0.29)	<u>68.84</u> (±0.16)	**70.28** (±0.10)
LED$_m$-G	Acc.	55.79	51.26	64.33 (±0.00)	37.75 (±0.06)	51.09 (±0.70)	55.54 (±1.00)	57.82 (±0.22)	<u>68.53</u> (±0.01)	**69.92** (±0.06)
	AUROC	44.23	54.76	69.75 (±0.00)	47.67 (±0.01)	51.52 (±1.57)	57.99 (±1.51)	60.95 (±0.22)	<u>75.22</u> (±0.11)	**77.35** (±0.04)
	AUPRC	58.26	55.98	69.59 (±0.00)	48.88 (±0.01)	51.94 (±1.48)	58.59 (±1.79)	60.78 (±0.10)	<u>74.76</u> (±0.15)	**77.28** (±0.03)
	Bal. Acc.	55.77	50.59	64.29 (±0.00)	37.83 (±0.06)	50.45 (±0.83)	55.43 (±1.03)	57.81 (±0.22)	<u>68.50</u> (±0.01)	**69.91** (±0.06)
SEA$_m$-A	Acc.	52.24	60.42	67.47 (±0.00)	52.99 (±0.09)	60.71 (±0.43)	66.96 (±0.34)	69.39 (±0.31)	<u>72.77</u> (±0.17)	**73.56** (±0.03)
	AUROC	49.01	68.89	73.07 (±0.00)	46.75 (±0.13)	62.64 (±5.43)	72.35 (±0.35)	74.09 (±0.64)	<u>78.55</u> (±0.14)	**79.20** (±0.02)
	AUPRC	68.43	78.00	79.60 (±0.00)	57.65 (±0.11)	73.02 (±3.19)	78.97 (±0.09)	79.72 (±0.47)	<u>82.57</u> (±0.09)	**83.18** (±0.04)
	Bal. Acc.	50.99	51.37	63.45 (±0.00)	44.89 (±0.11)	50.85 (±1.08)	63.20 (±0.33)	65.95 (±0.36)	<u>69.68</u> (±0.17)	**70.82** (±0.10)
SEA$_m$-G	Acc.	51.76	60.57	67.19 (±0.00)	53.11 (±0.13)	60.69 (±0.42)	66.73 (±0.41)	69.09 (±0.21)	<u>72.10</u> (±0.20)	**72.88** (±0.08)
	AUROC	49.60	68.15	72.38 (±0.00)	45.97 (±0.07)	62.21 (±5.60)	71.83 (±0.40)	73.52 (±0.18)	<u>77.46</u> (±0.13)	**78.22** (±0.01)
	AUPRC	68.77	77.25	78.89 (±0.00)	57.03 (±0.07)	72.54 (±3.20)	78.47 (±0.20)	79.11 (±0.16)	<u>81.71</u> (±0.11)	**82.36** (±0.01)
	Bal. Acc.	50.40	51.15	63.03 (±0.00)	44.99 (±0.10)	50.73 (±0.94)	62.70 (±0.33)	65.48 (±0.15)	<u>68.91</u> (±0.22)	**70.10** (±0.04)
STAGGER$_m$-A	Acc.	68.10	64.42	77.40 (±0.00)	60.13 (±0.04)	70.08 (±0.54)	75.19 (±0.67)	82.39 (±0.21)	**84.51** (±0.05)	<u>83.81</u> (±0.03)
	AUROC	33.10	36.75	86.05 (±0.00)	44.43 (±0.19)	69.41 (±2.95)	79.03 (±0.54)	91.79 (±0.10)	<u>92.62</u> (±0.06)	**93.11** (±0.04)
	AUPRC	44.95	31.49	79.63 (±0.00)	35.69 (±0.14)	56.20 (±5.39)	70.08 (±0.09)	89.28 (±0.15)	<u>90.69</u> (±0.10)	**91.23** (±0.05)
	Bal. Acc.	66.90	57.46	76.41 (±0.00)	50.32 (±0.05)	70.81 (±0.80)	73.54 (±0.62)	81.49 (±0.20)	<u>82.74</u> (±0.04)	**83.11** (±0.03)
STAGGER$_m$-G	Acc.	67.25	64.38	77.04 (±0.00)	60.04 (±0.02)	68.92 (±0.19)	74.93 (±0.73)	80.67 (±0.05)	**82.22** (±0.07)	<u>81.73</u> (±0.09)
	AUROC	34.20	37.16	85.50 (±0.00)	46.19 (±0.10)	67.91 (±3.36)	79.55 (±0.82)	89.48 (±0.10)	<u>90.56</u> (±0.07)	**91.08** (±0.04)
	AUPRC	46.33	31.81	79.30 (±0.00)	37.08 (±0.08)	57.00 (±1.50)	72.65 (±1.31)	86.42 (±0.14)	<u>87.96</u> (±0.09)	**88.63** (±0.04)
	Bal. Acc.	65.80	57.49	75.94 (±0.00)	50.30 (±0.02)	67.68 (±1.18)	73.15 (±0.64)	79.35 (±0.04)	<u>80.32</u> (±0.08)	**81.05** (±0.04)

Table 12: Detailed results on large-scale streams ($n \geq 10{,}000$) at $p = 0.25$. Higher is better for all metrics. Bold and underlined values indicate the best and second-best results, respectively.

Dataset	Metric	Classical Online Learning				Online Deep Learning				
		OLIFL	OLVF	OLWDF	ORF^{3}V	HBP	Aux-Drop	MODL	HI^2	Our
Real-world datasets										
phishing	Acc.	58.24	72.37	77.01 (±0.00)	51.04 (±0.20)	62.60 (±1.35)	56.78 (±1.37)	74.46 (±0.32)	74.60 (±0.73)	**80.81** (±0.12)
	AUROC	46.68	58.51	84.84 (±0.00)	45.86 (±0.02)	70.35 (±2.36)	60.62 (±3.54)	82.68 (±0.17)	82.21 (±0.72)	**89.00** (±0.06)
	AUPRC	43.24	62.79	86.56 (±0.00)	52.81 (±0.05)	75.33 (±1.85)	67.82 (±2.83)	85.43 (±0.12)	84.30 (±0.56)	**90.59** (±0.08)
	Bal. Acc.	53.18	70.94	76.23 (±0.00)	46.16 (±0.18)	58.81 (±1.69)	51.39 (±1.74)	73.59 (±0.32)	73.77 (±0.75)	**80.39** (±0.10)
weather	Acc.	65.79	61.80	68.24 (±0.35)	68.28 (±0.05)	68.54 (±0.04)	68.55 (±0.06)	68.58 (±0.01)	68.50 (±0.26)	**69.29** (±0.15)
	AUROC	49.07	48.29	60.78 (±0.86)	46.79 (±0.10)	54.43 (±0.73)	45.92 (±0.11)	50.00 (±0.01)	65.39 (±1.23)	**68.55** (±0.07)
	AUPRC	**61.69**	30.07	40.17 (±0.99)	28.98 (±0.06)	34.95 (±0.63)	28.90 (±0.32)	31.33 (±1.04)	43.42 (±1.21)	47.20 (±0.16)
	Bal. Acc.	50.92	54.34	52.71 (±0.23)	49.90 (±0.02)	50.01 (±0.03)	50.00 (±0.01)	50.01 (±0.01)	54.16 (±0.76)	**58.07** (±0.45)
magic04	Acc.	57.22	59.59	66.02 (±0.00)	37.96 (±0.19)	68.59 (±0.09)	67.63 (±0.26)	69.03 (±0.11)	69.70 (±0.30)	**70.80** (±0.10)
	AUROC	46.93	44.02	60.70 (±0.00)	48.00 (±0.09)	60.85 (±0.39)	58.07 (±0.75)	62.27 (±0.21)	66.23 (±0.68)	**68.78** (±0.15)
	AUPRC	68.63	59.34	72.48 (±0.00)	63.04 (±0.05)	70.25 (±0.28)	68.52 (±0.53)	71.34 (±0.29)	76.59 (±0.73)	**78.57** (±0.09)
	Bal. Acc.	53.06	53.18	54.30 (±0.00)	47.94 (±0.22)	57.35 (±0.10)	56.07 (±0.49)	59.52 (±0.16)	59.74 (±0.46)	**61.61** (±0.14)
a8a	Acc.	75.78	72.19	**78.44** (±0.00)	75.90 (±0.00)	76.71 (±0.36)	75.85 (±0.07)	76.02 (±0.06)	77.10 (±0.20)	78.39 (±0.09)
	AUROC	46.27	63.25	78.58 (±0.00)	47.88 (±0.05)	68.43 (±3.00)	55.92 (±3.39)	74.45 (±0.34)	73.95 (±0.28)	**80.12** (±0.07)
	AUPRC	66.67	85.07	91.75 (±0.00)	74.49 (±0.03)	87.26 (±1.49)	81.27 (±2.08)	89.62 (±0.22)	89.32 (±0.09)	**92.34** (±0.04)
	Bal. Acc.	53.57	60.67	60.70 (±0.00)	49.99 (±0.00)	53.15 (±1.38)	50.00 (±0.01)	50.29 (±0.15)	57.23 (±0.59)	**67.08** (±0.09)
elec	Acc.	54.65	55.18	64.70 (±0.20)	47.92 (±0.10)	59.78 (±0.44)	57.91 (±0.17)	63.58 (±0.14)	67.60 (±0.95)	**69.66** (±0.49)
	AUROC	47.92	51.46	69.62 (±0.38)	47.38 (±0.12)	60.48 (±0.57)	54.98 (±0.54)	66.58 (±0.21)	73.21 (±1.40)	**76.49** (±0.32)
	AUPRC	60.43	42.93	60.78 (±0.32)	40.81 (±0.06)	52.51 (±0.63)	47.04 (±0.61)	59.71 (±0.19)	66.26 (±2.21)	**71.41** (±0.28)
	Bal. Acc.	52.06	50.43	62.17 (±0.32)	47.52 (±0.09)	54.26 (±0.67)	50.86 (±0.42)	60.41 (±0.18)	65.93 (±0.98)	**68.29** (±0.53)
Synthetic datasets										
LED$_m$-A	Acc.	52.70	51.59	59.12 (±0.00)	41.42 (±0.03)	50.86 (±0.29)	52.03 (±0.21)	54.04 (±0.23)	60.18 (±0.40)	**62.06** (±0.02)
	AUROC	47.32	52.73	63.32 (±0.00)	48.62 (±0.01)	50.77 (±0.65)	53.29 (±0.36)	55.95 (±0.42)	64.83 (±0.32)	**67.69** (±0.04)
	AUPRC	60.78	53.99	62.84 (±0.00)	49.57 (±0.02)	51.62 (±0.87)	54.19 (±0.42)	56.93 (±0.25)	64.43 (±0.26)	**67.78** (±0.04)
	Bal. Acc.	52.68	50.99	59.05 (±0.00)	41.44 (±0.03)	50.29 (±0.60)	51.72 (±0.28)	54.05 (±0.23)	60.10 (±0.40)	**62.03** (±0.01)
LED$_m$-G	Acc.	52.33	51.50	59.11 (±0.00)	41.64 (±0.06)	50.79 (±0.20)	51.76 (±0.17)	53.74 (±0.27)	60.12 (±0.35)	**62.19** (±0.08)
	AUROC	47.70	52.64	63.37 (±0.00)	48.72 (±0.01)	50.77 (±0.41)	52.85 (±0.21)	55.60 (±0.18)	64.51 (±0.35)	**67.45** (±0.06)
	AUPRC	60.76	53.74	62.97 (±0.00)	49.60 (±0.02)	51.44 (±0.66)	53.59 (±0.24)	56.15 (±0.11)	64.11 (±0.24)	**67.47** (±0.06)
	Bal. Acc.	52.30	50.94	59.04 (±0.00)	41.67 (±0.06)	50.24 (±0.49)	51.44 (±0.21)	53.73 (±0.27)	60.05 (±0.35)	**62.15** (±0.07)
SEA$_m$-A	Acc.	53.18	60.46	65.82 (±0.00)	50.68 (±0.14)	60.36 (±0.05)	65.25 (±0.12)	66.95 (±0.17)	68.13 (±0.12)	**68.64** (±0.05)
	AUROC	48.47	67.03	70.11 (±0.00)	45.82 (±0.10)	60.33 (±5.36)	69.31 (±0.08)	70.80 (±0.11)	72.69 (±0.16)	**73.61** (±0.01)
	AUPRC	67.62	75.22	77.56 (±0.00)	57.92 (±0.05)	71.51 (±3.00)	76.26 (±0.16)	77.28 (±0.11)	79.03 (±0.16)	**79.69** (±0.03)
	Bal. Acc.	51.53	50.00	61.38 (±0.00)	43.40 (±0.14)	50.05 (±0.17)	60.91 (±0.25)	62.51 (±0.23)	64.17 (±0.16)	**65.09** (±0.04)
SEA$_m$-G	Acc.	52.84	60.46	65.47 (±0.00)	50.92 (±0.17)	60.36 (±0.04)	65.25 (±0.25)	66.89 (±0.07)	68.07 (±0.14)	**68.85** (±0.01)
	AUROC	48.96	66.56	69.51 (±0.00)	45.28 (±0.11)	59.85 (±5.46)	68.93 (±0.11)	70.23 (±0.36)	72.18 (±0.15)	**73.16** (±0.01)
	AUPRC	67.85	74.64	77.21 (±0.00)	57.40 (±0.13)	71.19 (±2.92)	76.21 (±0.07)	76.92 (±0.31)	78.67 (±0.11)	**79.41** (±0.01)
	Bal. Acc.	51.04	50.00	60.78 (±0.00)	43.64 (±0.15)	50.05 (±0.16)	60.51 (±0.26)	62.53 (±0.43)	64.13 (±0.17)	**65.41** (±0.04)
STAGGER$_m$-A	Acc.	68.48	60.57	75.53 (±0.00)	60.18 (±0.04)	69.12 (±1.19)	73.82 (±0.67)	78.67 (±0.14)	**79.39** (±0.09)	79.29 (±0.28)
	AUROC	32.54	44.88	83.99 (±0.00)	45.18 (±0.07)	63.38 (±7.45)	77.50 (±1.05)	87.84 (±0.11)	88.24 (±0.07)	**88.79** (±0.07)
	AUPRC	44.20	36.17	76.27 (±0.00)	35.51 (±0.04)	51.49 (±10.58)	67.82 (±1.26)	84.11 (±0.19)	84.95 (±0.13)	**85.74** (±0.07)
	Bal. Acc.	67.46	51.07	74.92 (±0.00)	50.42 (±0.03)	69.40 (±2.92)	72.15 (±0.66)	77.71 (±0.13)	77.22 (±0.10)	**78.75** (±0.24)
STAGGER$_m$-G	Acc.	67.32	60.32	74.24 (±0.00)	60.03 (±0.04)	68.27 (±1.49)	73.48 (±0.38)	77.02 (±0.18)	**77.80** (±0.07)	77.66 (±0.15)
	AUROC	33.99	44.88	82.73 (±0.00)	46.49 (±0.12)	60.68 (±7.13)	77.73 (±0.61)	85.78 (±0.10)	86.28 (±0.11)	**86.92** (±0.02)
	AUPRC	45.86	36.34	75.03 (±0.00)	36.66 (±0.08)	49.70 (±9.91)	69.15 (±1.03)	81.68 (±0.12)	82.22 (±0.25)	**83.31** (±0.03)
	Bal. Acc.	66.01	50.76	73.64 (±0.00)	50.33 (±0.04)	67.19 (±3.29)	71.87 (±0.31)	75.36 (±0.12)	75.52 (±0.15)	**77.12** (±0.04)

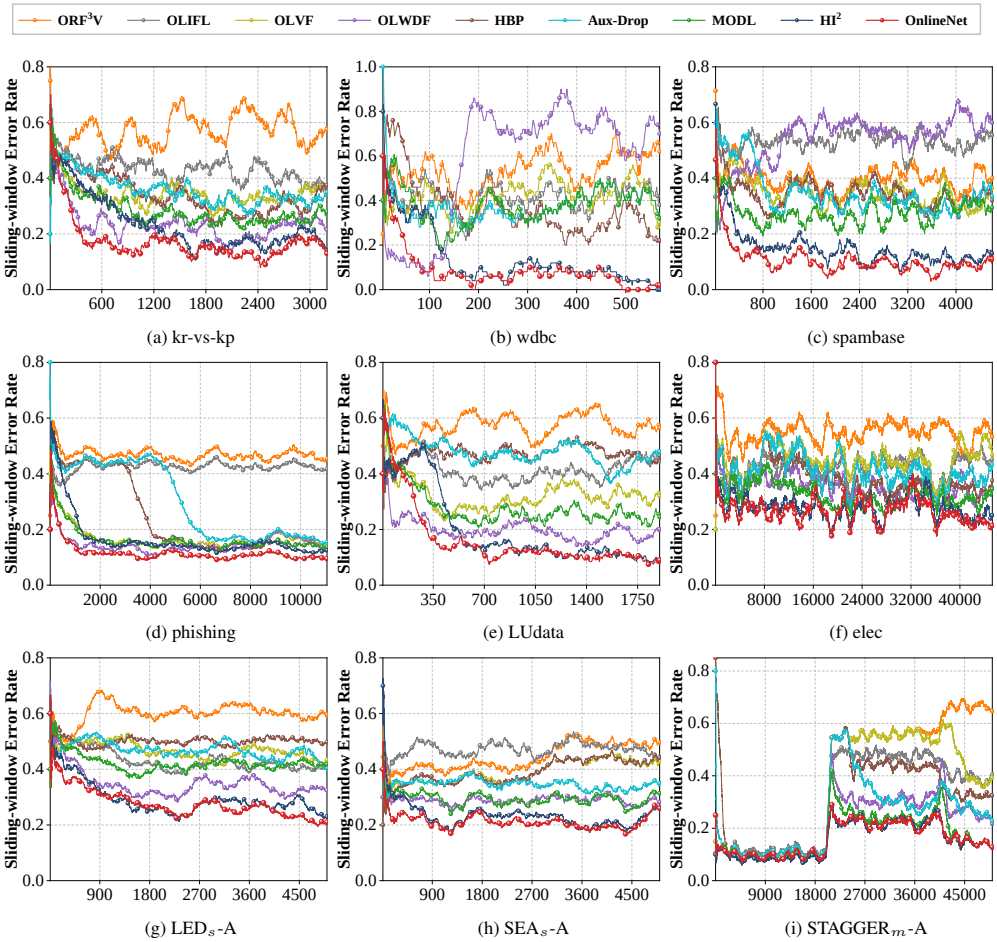


Figure 6: Performance evolution of different methods as instances arrive sequentially at $p = 0.50$.

E.2 Supplementary Results on Online Performance Evolution

Figures 6 and 7 present the corresponding results under lower feature availability at $p = 0.50$ and $p = 0.25$. Compared with the results at $p = 0.75$ in Figure 1, lower feature availability generally leads to higher local error levels, while the extent of degradation varies across datasets and methods. At $p = 0.50$, OnlineNet maintains low error levels across the streams. HI^2 and OLWDF occasionally approach its performance over limited intervals, while the overall performance pattern remains broadly similar to that observed at $p = 0.75$. Under the more challenging $p = 0.25$ setting, error levels and fluctuations increase on several streams, yet OnlineNet continues to maintain comparatively low error levels. On STAGGER$_m$-A, the effects of concept transitions are more apparent. OnlineNet shows a comparatively limited increase in error following the initial transition, while HI^2 remains competitive and the performance differences among the leading methods decrease after the later transition. Overall, these results indicate that OnlineNet preserves strong local predictive performance as feature availability decreases and remains robust as the data stream evolves.

E.3 Supplementary Complexity Analysis

Time Complexity Let d denote the input dimension, h the hidden dimension, L the number of paired fast–slow layers, and C the number of output classes. For each arriving instance, causal normalization and the update of feature-wise running statistics require $\mathcal{O}(d)$ operations. The slow pathway costs $\mathcal{O}(dh + (L-1)h^2)$. The fast pathway costs $\mathcal{O}(2dh + (L-1)h^2)$ because its first layer receives the concatenation of the zero-filled input and its availability mask. BAM adds $\mathcal{O}(Lh)$

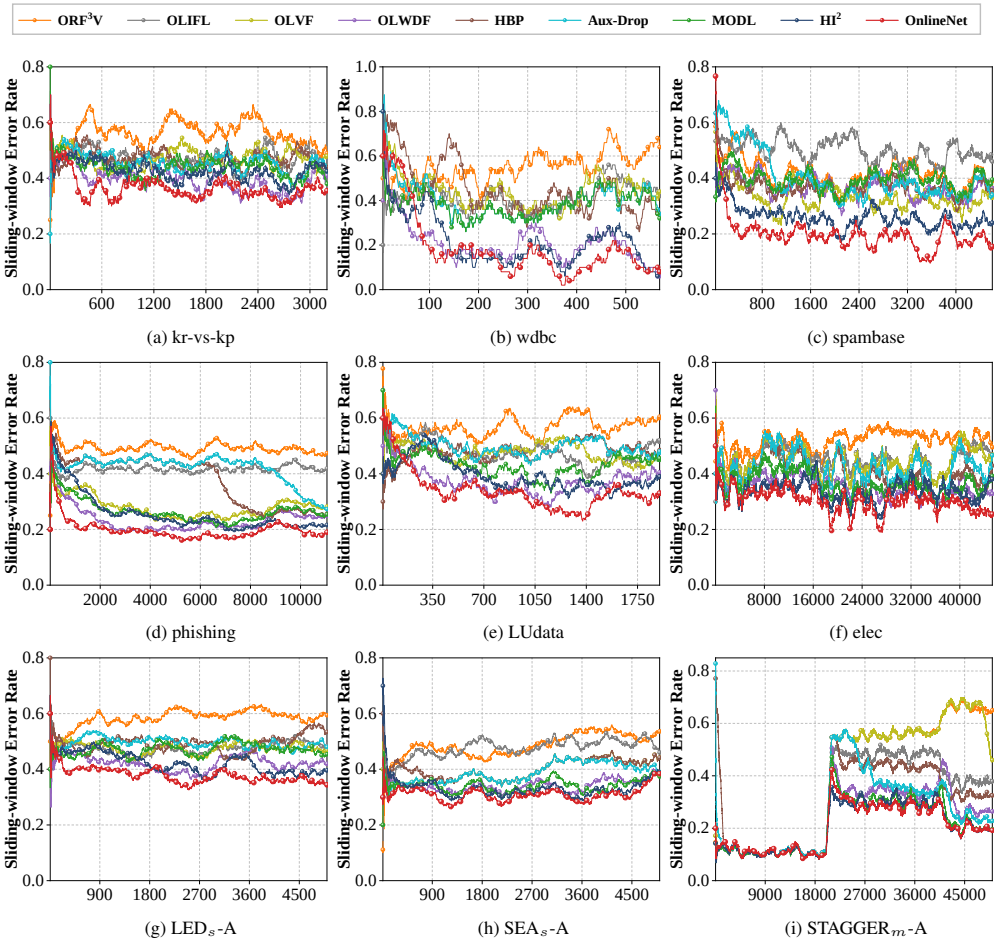


Figure 7: Performance evolution of different methods as instances arrive sequentially at $p = 0.25$.

element-wise operations, and the output classifier costs $\mathcal{O}(Ch)$. Computing the prequential class prior and applying the logit adjustment require a further $\mathcal{O}(C)$ operations. The complexity of one forward evaluation is therefore

$$\mathcal{O}\left(d + 3dh + 2(L-1)h^2 + Lh + Ch + C\right) = \mathcal{O}(dh + Lh^2 + Ch). \tag{20}$$

The test-then-train procedure evaluates the network once for prediction and again for training. The backward pass and AdamW update have the same asymptotic order as the training forward evaluation because they scale with the effective number of trainable parameters. The additional forward evaluation changes only the constant factor, so the per-instance complexity remains $\mathcal{O}(dh + Lh^2 + Ch)$. Processing a stream of T instances requires $\mathcal{O}(T(dh + Lh^2 + Ch))$ time. The per-instance cost is independent of the number of previously observed instances because OnlineNet does not replay historical instances.

Space Complexity Ignoring bias terms, the effective trainable parameter count of the paired pathways and output classifier is

$$P_{\text{eff}} = 3dh + 2(L-1)h^2 + Ch = \mathcal{O}(dh + Lh^2 + Ch). \tag{21}$$

Causal normalization maintains feature-wise counts, means, and second-order statistics in $\mathcal{O}(d)$ space. The prequential adjustment stores class counts in $\mathcal{O}(C)$ space. Temporary activations for one instance require $\mathcal{O}(d + Lh + C)$ memory. Gradients and the first- and second-moment states maintained by AdamW each scale as $\mathcal{O}(P_{\text{eff}})$. Hence, the overall training-space complexity is

$$\mathcal{O}\left(P_{\text{eff}} + d + Lh + C\right) = \mathcal{O}\left(dh + Lh^2 + Ch + d + C\right). \tag{22}$$

Table 13: Per-dataset ablation results on six datasets at $p = 0.75$. Values are reported as mean $\pm$ standard deviation.

Dataset	Metric	w/o Causal	w/o Fast$_{v1}$	w/o Fast$_{v2}$	w/o BAM$_{v1}$	w/o BAM$_{v2}$	w/o OPA	OnlineNet
a8a	Accuracy	82.53 $\pm$ 0.10	83.01 $\pm$ 0.07	82.95 $\pm$ 0.06	82.96 $\pm$ 0.08	82.92 $\pm$ 0.06	83.32 $\pm$ 0.06	82.93 $\pm$ 0.03
	AUROC	87.55 $\pm$ 0.04	88.05 $\pm$ 0.04	87.96 $\pm$ 0.03	87.84 $\pm$ 0.05	87.84 $\pm$ 0.07	87.94 $\pm$ 0.03	87.96 $\pm$ 0.03
	AUPRC	95.66 $\pm$ 0.02	95.80 $\pm$ 0.02	95.75 $\pm$ 0.02	95.72 $\pm$ 0.02	95.74 $\pm$ 0.02	95.73 $\pm$ 0.02	95.75 $\pm$ 0.02
	Bal. Acc.	75.67 $\pm$ 0.14	76.63 $\pm$ 0.07	76.50 $\pm$ 0.05	76.13 $\pm$ 0.16	76.12 $\pm$ 0.09	74.18 $\pm$ 0.09	76.65 $\pm$ 0.08
elec	Accuracy	72.75 $\pm$ 0.51	76.36 $\pm$ 0.13	76.28 $\pm$ 0.32	78.62 $\pm$ 0.59	78.44 $\pm$ 0.46	77.99 $\pm$ 0.40	77.72 $\pm$ 0.45
	AUROC	79.46 $\pm$ 0.52	84.83 $\pm$ 0.10	84.78 $\pm$ 0.28	86.94 $\pm$ 0.52	86.67 $\pm$ 0.34	86.57 $\pm$ 0.35	86.40 $\pm$ 0.41
	AUPRC	75.45 $\pm$ 0.72	81.37 $\pm$ 0.14	81.33 $\pm$ 0.32	83.95 $\pm$ 0.52	83.46 $\pm$ 0.38	83.11 $\pm$ 0.36	83.00 $\pm$ 0.51
	Bal. Acc.	71.21 $\pm$ 0.56	75.67 $\pm$ 0.09	75.55 $\pm$ 0.37	77.83 $\pm$ 0.65	77.79 $\pm$ 0.47	77.18 $\pm$ 0.45	77.15 $\pm$ 0.43
german	Accuracy	65.86 $\pm$ 1.08	72.54 $\pm$ 0.73	72.28 $\pm$ 1.08	68.66 $\pm$ 0.52	68.66 $\pm$ 1.00	72.88 $\pm$ 0.45	72.72 $\pm$ 0.54
	AUROC	57.90 $\pm$ 1.15	72.29 $\pm$ 0.56	71.98 $\pm$ 0.34	57.40 $\pm$ 2.08	62.96 $\pm$ 1.35	71.50 $\pm$ 0.77	71.70 $\pm$ 0.60
	AUPRC	73.90 $\pm$ 0.67	85.03 $\pm$ 0.25	84.73 $\pm$ 0.32	74.67 $\pm$ 1.34	76.63 $\pm$ 0.68	84.46 $\pm$ 0.59	84.51 $\pm$ 0.52
	Bal. Acc.	54.60 $\pm$ 1.10	62.52 $\pm$ 0.75	62.37 $\pm$ 1.21	52.64 $\pm$ 1.75	57.94 $\pm$ 1.40	60.97 $\pm$ 0.95	63.37 $\pm$ 0.42
phishing	Accuracy	90.77 $\pm$ 0.10	92.33 $\pm$ 0.08	92.25 $\pm$ 0.13	91.99 $\pm$ 0.12	92.19 $\pm$ 0.05	92.29 $\pm$ 0.12	92.31 $\pm$ 0.12
	AUROC	97.08 $\pm$ 0.01	97.94 $\pm$ 0.02	97.92 $\pm$ 0.03	97.67 $\pm$ 0.01	97.85 $\pm$ 0.03	97.90 $\pm$ 0.04	97.91 $\pm$ 0.03
	AUPRC	97.64 $\pm$ 0.01	98.32 $\pm$ 0.02	98.29 $\pm$ 0.03	98.07 $\pm$ 0.03	98.24 $\pm$ 0.03	98.27 $\pm$ 0.03	98.28 $\pm$ 0.04
	Bal. Acc.	90.61 $\pm$ 0.10	92.21 $\pm$ 0.08	92.13 $\pm$ 0.15	91.86 $\pm$ 0.12	92.07 $\pm$ 0.05	92.13 $\pm$ 0.13	92.18 $\pm$ 0.13
spam	Accuracy	95.30 $\pm$ 0.18	95.39 $\pm$ 0.22	95.36 $\pm$ 0.15	94.60 $\pm$ 0.55	95.45 $\pm$ 0.17	95.70 $\pm$ 0.11	95.61 $\pm$ 0.06
	AUROC	98.32 $\pm$ 0.06	98.65 $\pm$ 0.06	98.61 $\pm$ 0.05	98.08 $\pm$ 0.12	98.63 $\pm$ 0.03	98.72 $\pm$ 0.03	98.70 $\pm$ 0.06
	AUPRC	96.05 $\pm$ 0.17	96.89 $\pm$ 0.10	96.85 $\pm$ 0.11	95.77 $\pm$ 0.24	96.52 $\pm$ 0.11	97.00 $\pm$ 0.08	96.94 $\pm$ 0.10
	Bal. Acc.	94.60 $\pm$ 0.31	94.93 $\pm$ 0.34	94.83 $\pm$ 0.28	94.36 $\pm$ 0.57	94.91 $\pm$ 0.27	95.35 $\pm$ 0.19	95.16 $\pm$ 0.09
svmguide3	Accuracy	76.64 $\pm$ 0.75	78.21 $\pm$ 0.70	78.54 $\pm$ 0.42	77.94 $\pm$ 0.54	78.33 $\pm$ 0.54	78.00 $\pm$ 0.58	78.65 $\pm$ 0.55
	AUROC	60.96 $\pm$ 1.17	68.48 $\pm$ 1.39	68.31 $\pm$ 1.21	64.04 $\pm$ 1.88	65.75 $\pm$ 1.76	68.06 $\pm$ 1.31	68.60 $\pm$ 1.34
	AUPRC	80.65 $\pm$ 0.58	85.42 $\pm$ 0.92	85.37 $\pm$ 0.71	82.27 $\pm$ 1.15	83.22 $\pm$ 1.20	85.18 $\pm$ 0.80	85.38 $\pm$ 0.73
	Bal. Acc.	53.87 $\pm$ 1.40	58.67 $\pm$ 1.58	59.14 $\pm$ 1.34	57.35 $\pm$ 1.34	58.63 $\pm$ 1.40	56.74 $\pm$ 1.22	60.02 $\pm$ 1.25

OnlineNet does not retain past instances, so its online model state is independent of the stream length T. This algorithmic bound excludes storage of the externally supplied stream and evaluation records.

E.4 Supplementary Ablation Studies

Table 13 reports the per-dataset results. OnlineNet does not lead every dataset–metric pair, which is reasonable given the variation in data characteristics and realized availability patterns across the six streams. A simplified variant may align better with a particular dataset or evaluation metric; for example, **w/o BAM**$_{v1}$ performs best on elec, whereas **w/o OPA** is favored on spam. Several other reversals are small relative to the reported run-to-run variation. We therefore interpret the ablation through the cross-dataset averages in Table 3, rather than requiring every ablation to degrade every metric on every dataset. OnlineNet achieves the best average AUROC, AUPRC, and balanced accuracy, while **w/o OPA** yields only a marginal improvement in average accuracy. These results support the complete design at the aggregate level, while showing that the contribution of an individual component can depend on the dataset.

E.5 Supplementary Hyperparameter Analysis

This section supplements the experimental results and analysis regarding hyperparameters (i.e., the modulation coefficient and the adjustment coefficient) using a broader set of datasets.

E.5.1 Supplementary Results on Modulation Coefficient

Figure 8 reports the effect of the modulation coefficient on balanced accuracy across additional datasets. Values of $\lambda < 0$ generally yield lower balanced accuracy and higher variance, consistent with the findings in Section 5.4.1.

E.5.2 Supplementary Results on Adjustment Coefficient

Figure 9 compares overall accuracy and balanced accuracy across six additional datasets. The effect of τ depends on the stream. Stronger adjustment produces clear balanced-accuracy gains on

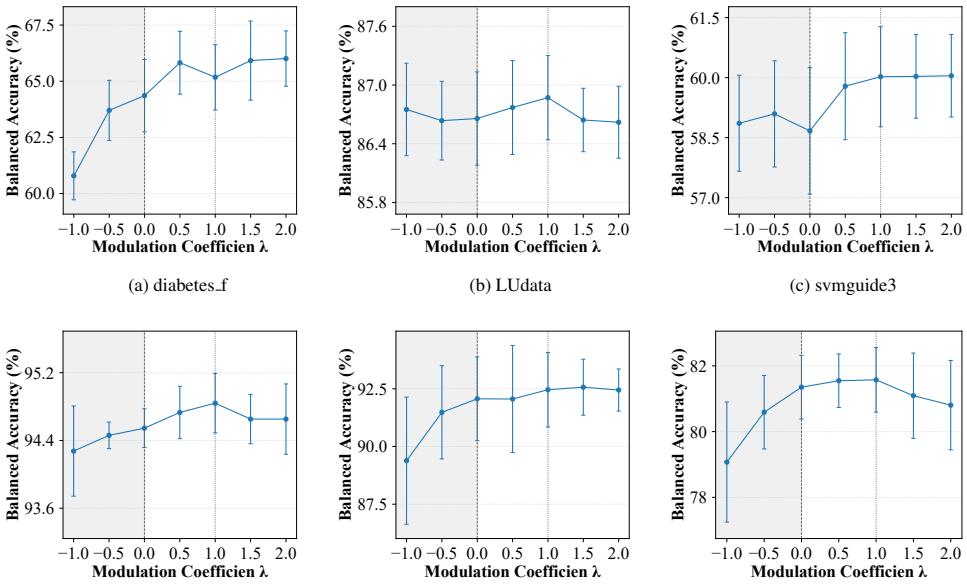


Figure 8: Effect of modulation coefficient λ under different datasets at $p = 0.75$. Error bars indicate standard deviations.

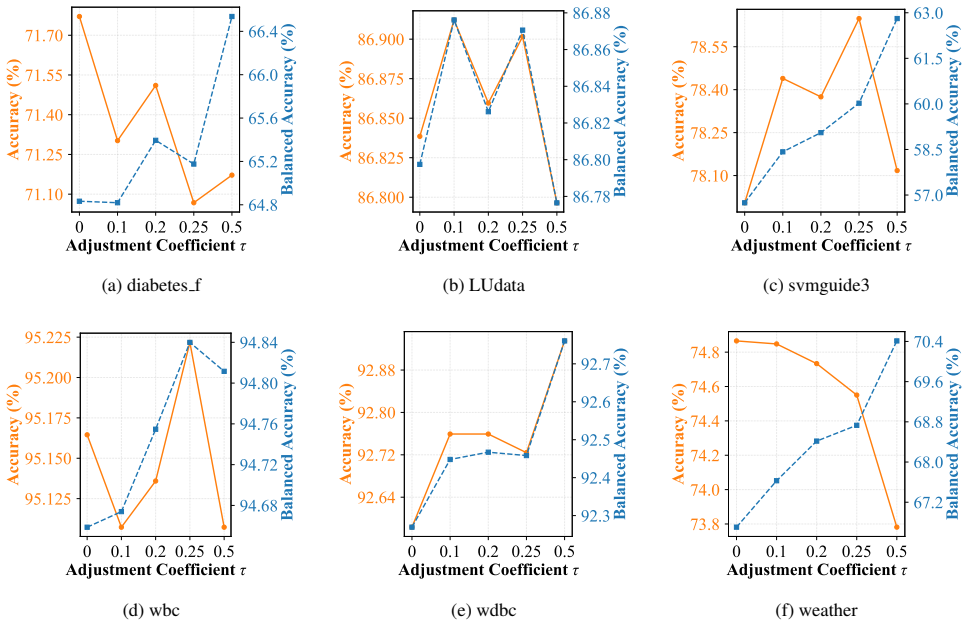


Figure 9: Effect of adjustment coefficient τ under different datasets at $p = 0.75$.

svmguide3, weather, and diabetes_f, whereas the changes on LUdata, wbc, and wdbc are smaller or non-monotonic. Accordingly, $\tau = 0.25$ is a reasonable default when no stream-specific tuning set is available. A stronger adjustment can be considered for markedly imbalanced streams when balanced accuracy is the primary criterion, provided that the accompanying change in overall accuracy is acceptable. For streams with weak or non-monotonic sensitivity, a smaller coefficient should be preferred to avoid unnecessary correction.

F Additional Exploratory Analyses

This section presents two exploratory analyses beyond the main experiments. We investigate how learning rate choices affect an online neural network with fast–slow pathways and evaluate the potential of Transformer-based architectures under the same strict online protocol.

F.1 Learning-Rate Sensitivity of the Fast and Slow Pathways

The proposed fast–slow design suggests using a higher learning rate for the fast pathway and a lower rate for the slow pathway. We examine this intuition by independently varying the learning rates of the two pathways over $\{10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$ at $p = 0.75$. Figure 10 shows that performance is generally more sensitive to the learning rate of the slow pathway. On a8a, reducing the learning rate of the slow pathway improves all three metrics, while changing the learning rate of the fast pathway at a fixed slow-pathway rate has only a limited effect. On elec and svmguide3, the best or tied-best results generally occur when both pathways use 10^{-3}. diabetes_f is the main exception, where using 10^{-3} for the slow pathway and 10^{-2} for the fast pathway yields the highest performance across the three metrics. Asymmetric learning rates provide no improvement on elec and no visible gain on svmguide3 at the reported precision.

These results do not support a general rule that the fast pathway should use a larger learning rate. In OnlineNet, the terms fast and slow describe how the pathways use stream information. They do not imply that their parameters must be updated at different rates. The fast pathway can remain responsive to the current observation because it receives the current values directly and modulates the slow representation through BAM, even when the two pathways share a learning rate. Given the small and dataset-dependent gains from separate rates, we use a single rate for the two pathways in our experiments.

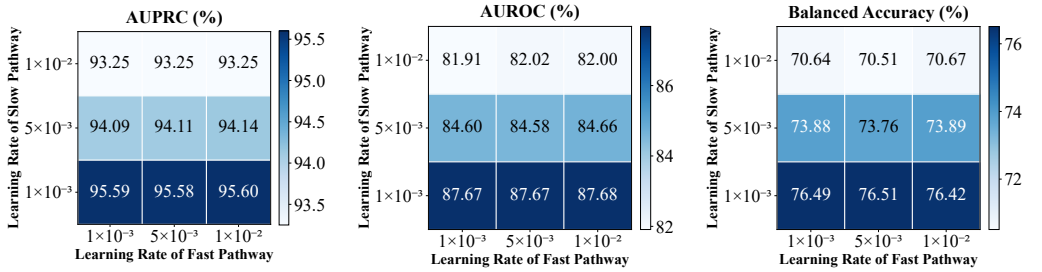
F.2 Exploring Transformer-Based Online Architectures

We explore five Transformer-based architectures under the strict online protocol considered in this work. Each model first predicts on the current instance, performs a single supervised model update after the label is revealed, and then discards the instance without subsequent replay or reuse.

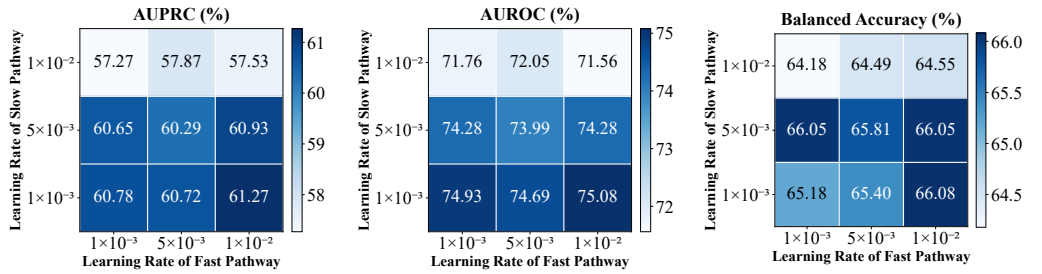
Architectures We consider three compact Transformer Vaswani et al. (2017) variants with a common encoder configuration of six layers, a hidden dimension of 64, and four attention heads per layer. For each instance, each observed feature is represented by its feature identifier and numerical value, while unavailable features are omitted. **Vanilla Transformer (V-Transformer)** mean-pools the final-layer representations and applies a classification head. **Online Adaptive Transformer (OA-Transformer)** additionally attaches prediction heads to intermediate layers and aggregates their outputs using online Hedge weights. **SMoD-ORA Transformer (SO-Transformer)** further introduces token-level gates on the residual updates together with an orthogonality regularization term between the input representation and the residual update.

We further evaluate two alternative Transformer-based designs that explore different representation paradigms. **BERT-Online** Devlin et al. (2019) maps each numerical feature to a token, represents unavailable features with a learned embedding, and feeds the resulting sequence to a two-layer BERT encoder initialized from scratch for classification. **Qwen–LoRA** Qwen et al. (2025) uses a pretrained Qwen2.5-1.5B model for sequence classification, with the observed feature–value pairs serialized as text. Online adaptation is performed using rank-8 LoRA Hu et al. (2021) applied to the query and value projections.

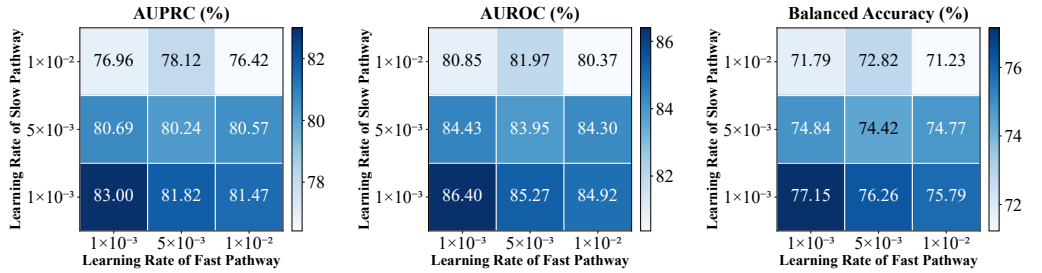
Discussion Table 14 summarizes the comparison between OnlineNet and five Transformer-based online learners on nine datasets evaluated by all models. OnlineNet achieves the highest mean balanced accuracy among the evaluated models, reaching 79.76%. Among the Transformer-based alternatives, Qwen–LoRA achieves the strongest overall performance, with a mean balanced accuracy of 64.26%. Within the compact encoder variants, OA-Transformer improves over V-Transformer by 6.85 percentage points, suggesting that intermediate-layer predictions combined with online aggregation can benefit adaptation under strict online updates. In contrast, V-Transformer and BERT-Online perform substantially worse, suggesting that directly using a Transformer backbone does



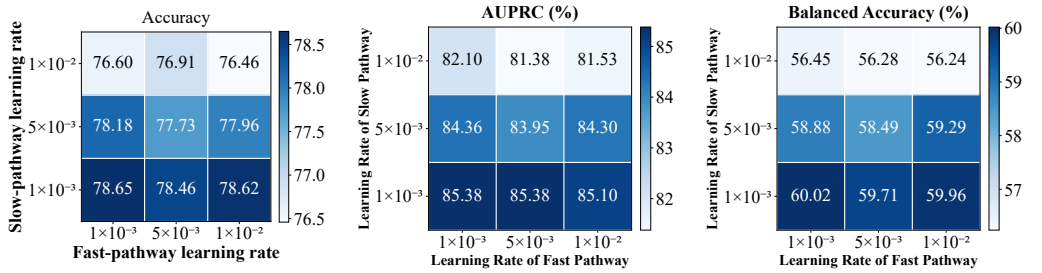
(a) a8a



(b) diabetes_f



(c) elec



(d) svmguide3

Figure 10: Sensitivity to the learning rates of the fast and slow pathways. Darker colors indicate better performance.

Table 14: Comparison results on nine datasets at $p = 0.75$. Bold and underlined values denote the best and second-best results, respectively. The last block reports the average performance across the nine datasets.

Dataset	Metric	V-Transformer	OA-Transformer	SO-Transformer	BERT-Online	Qwen–LoRA	OnlineNet
a8a	Accuracy	75.90	<u>78.95</u>	74.51	75.70	77.59	**82.93**
	AUROC	50.24	<u>79.85</u>	67.82	49.52	76.75	**87.96**
	AUPRC	76.17	<u>92.01</u>	85.71	75.76	91.13	**95.75**
	Bal. Acc.	50.00	<u>63.73</u>	57.76	50.05	61.86	**76.65**
australian	Accuracy	54.03	<u>60.29</u>	54.87	50.19	55.56	**81.97**
	AUROC	53.34	<u>62.53</u>	55.79	48.26	54.76	**87.84**
	AUPRC	46.35	<u>59.76</u>	56.47	43.88	48.01	**85.57**
	Bal. Acc.	52.18	<u>59.75</u>	53.80	48.68	53.73	**81.58**
diabetes_f	Accuracy	62.29	<u>63.02</u>	59.53	59.07	60.63	**71.07**
	AUROC	48.80	49.01	52.12	51.78	<u>55.54</u>	**74.93**
	AUPRC	33.41	35.80	<u>40.16</u>	36.29	38.69	**60.78**
	Bal. Acc.	49.05	51.03	<u>53.84</u>	51.20	50.95	**65.18**
krvskp	Accuracy	51.44	<u>62.24</u>	51.28	51.00	55.06	**83.37**
	AUROC	50.90	<u>66.26</u>	52.23	51.02	56.86	**91.64**
	AUPRC	52.10	<u>66.89</u>	55.04	52.93	57.81	**92.69**
	Bal. Acc.	50.80	<u>62.04</u>	51.08	50.74	54.77	**83.37**
magic04	Accuracy	64.65	64.58	60.97	63.22	<u>74.53</u>	**78.60**
	AUROC	50.29	53.34	56.38	49.97	<u>77.90</u>	**84.24**
	AUPRC	64.77	66.29	67.85	64.71	<u>84.95</u>	**89.64**
	Bal. Acc.	50.02	50.92	55.30	49.85	<u>68.89</u>	**74.48**
spambase	Accuracy	59.68	66.13	66.01	57.07	<u>74.69</u>	**89.90**
	AUROC	49.66	67.76	67.86	49.93	<u>81.57</u>	**95.02**
	AUPRC	39.40	58.56	59.43	39.98	<u>76.62</u>	**92.17**
	Bal. Acc.	50.29	61.13	62.80	50.10	<u>72.39</u>	**89.27**
svmguide3	Accuracy	<u>75.96</u>	75.86	73.92	73.02	74.82	**78.65**
	AUROC	50.12	<u>52.08</u>	50.37	48.52	51.72	**68.60**
	AUPRC	75.72	<u>76.42</u>	75.79	75.11	76.07	**85.38**
	Bal. Acc.	50.41	50.93	49.63	50.71	<u>50.96</u>	**60.02**
wbc	Accuracy	89.27	92.82	<u>93.85</u>	58.18	89.13	**95.22**
	AUROC	94.05	97.39	<u>97.57</u>	51.90	94.57	**98.39**
	AUPRC	87.21	93.85	<u>96.01</u>	35.34	89.81	**96.53**
	Bal. Acc.	87.57	91.96	<u>92.46</u>	50.49	86.56	**94.84**
wdbc	Accuracy	57.43	64.36	70.16	54.89	<u>80.37</u>	**92.72**
	AUROC	47.07	56.82	73.57	49.03	<u>88.06</u>	**97.24**
	AUPRC	35.13	49.87	71.57	36.84	<u>81.72</u>	**96.48**
	Bal. Acc.	47.38	57.90	68.65	49.04	<u>78.23</u>	**92.46**
Average	Accuracy	65.63	69.81	67.23	60.26	<u>71.38</u>	**83.83**
	AUROC	54.94	65.00	63.74	49.99	<u>70.86</u>	**87.32**
	AUPRC	56.70	66.61	67.56	51.20	<u>71.64</u>	**88.33**
	Bal. Acc.	54.19	61.04	60.59	50.10	<u>64.26</u>	**79.76**

not ensure effective single-instance adaptation under the evaluated protocol. Notably, Qwen–LoRA achieves this performance despite representing numerical feature–value pairs as text, suggesting that the pretrained Qwen model can extract useful information from serialized tabular representations. Future work could examine how such models capture numerical structure when feature availability changes across instances.

However, Transformer-based architectures incur substantial computational and memory overhead in strict online learning. Self-attention scales quadratically with the number of input tokens, and each arriving instance requires a separate forward and backward pass that cannot be readily amortized through batching. Although parameter-efficient methods such as LoRA reduce the number of trainable parameters, the full pretrained backbone must still be stored and executed at each update. These limitations may restrict their use in real-time streaming applications, particularly under resource constraints.