# Position Paper: The Significance of Class Imbalance in Online Semi-Supervised Data Stream Learning

Hadi Talal Jaafar Al-Kadhimi
*School of Computer Science, University of Birmingham*
Birmingham, United Kingdom
hxa568@student.bham.ac.uk

Leandro L. Minku
*School of Computer Science, University of Birmingham*
Birmingham, United Kingdom
l.l.minku@bham.ac.uk

***Abstract*—An important challenge in machine learning is learning from data streams in changing environments, especially when the data is imbalanced and only a few labels are available. A useful approach is semi-supervised online learning, which can take advantage of many unlabelled examples to help the model adapt over time. However, standard methods often have difficulty when imbalance occurs, since the majority class tends to dominate and the minority class is often ignored. In this paper, we show that the challenge of class imbalance is exacerbated by lack of labels in semi-supervised data stream learning, such that tackling class imbalance is an even more significant issue in this context than in supervised data stream learning. However, very little has been done in this area so far. Our study alerts the community of the need for novel semi-supervised strategies that can benefit from unlabelled examples to cope with class imbalance in streaming scenarios, rather than propagating the bias towards the majority class through these examples.**

***Index Terms*—Data stream learning, semi-supervised online learning, concept drift, class imbalance**

## I. Introduction

Data streams are continuous, possibly infinite sequences of instances generated in real time from sources like sensors, online services, financial transactions, and social media. Unlike static datasets,they cannot be completely stored or entirely available beforehand. So, learning algorithms must be able to process incoming examples that arrive over time [1].

Learning over time is particularly important as data streams may suffer concept drift, i.e., changes in the underlying distribution [2]. These changes can be a result of, e.g., seasonal behaviour, sensor recalibration, adversarial behaviour or other hidden contexts [1]. Concept drifts can make decision boundaries that had been learned earlier become less useful. Without learning new incoming examples from the new distribution, models risk becoming obsolete and underperforming. To avoid delays in adaptation, learning should ideally be in online mode, i.e., new incoming examples should be processed and learned as soon as they become available.

However, labelling incoming examples for training is an expensive and time consuming, as it typically requires manual effort. Even though real world applications may generate vast amounts of data over time, only a limited proportion can normally be labelled. For example, fraud vs genuine labels for credit card transactions depend on investigations and delayed feedback, typically available only for a subset of transactions [3]. In healthcare, annotating ECG data is also costly [4]. For example, labelling 15,000 short ECG records may require doctors working for almost three months. Similar challenges appear in intrusion detection, where creating labelled data requires expert knowledge and time [5]. The field of semi-supervised data stream learning has thus gained traction in recent years [6]–[12], which uses both unlabelled and labelled data for model adaptation.

Despite its promise, existing semi-supervised data stream learning overlook a significant issue: data streams are often imbalanced in real-life scenarios, with events (classes) of interest occurring infrequently but being extremely important. For example, in the EU/EEA, card-payment fraud accounts for around 0.015%-0.017% of all card transactions (in volume terms) [13]. Life-threatening cardiac abnormalities may only happen once every tens of thousands of heartbeats in intensive care monitoring, where failing to detect even one event could be fatal [14]. Similar imbalance patterns are also seen in autonomous systems, industrial fault monitoring, and cybersecurity intrusion detection.

It is well-known that class imbalance, if uncatered, can result in models that perform poorly on the minority class. In the worst case scenario, a learning algorithm that overlooks class imbalance could result in models that always predict the majority class, independent of the input features of the examples being predicted [15].

In this position paper, we argue that *class imbalance can be an even more challenging problem in semi-supervised data stream learning than when assuming all incoming examples arrive with labels. Therefore, it is crucial for the community to advance the field by tackling class imbalance in semi-supervised data stream learning.*

Our argument is based on: (1) a discussion of existing work, highlighting that existing approaches are unlikely sufficient to tackle class imbalance in semi-supervised data stream learning, and (2) an analysis of the impact of class imbalance on a state-of-the-art online semi-supervised neural network, showing that the deterioration of predictive performance in the presence of class imbalance can be even more severe when incoming examples have missing labels than when labels are present.

This paper is further organised as follows. Section II formu-

Hadi T. J. Al-Kadhimi acknowledges the full financial support provided by the Higher Committee for Education Development in Iraq (HCED) for his PhD studies.

lates the problem of online semi-supervised data stream learning in the presence of class imbalance. Section III discusses existing work. Section IV introduces the research questions answered by our experimental study and the methodology to answer these research questions. Section V analyses the results. Section VI concludes the paper. All data and code used in this study will be made available online.

## II. Problem Formulation

Let a data stream be a potentially infinite sequence $\mathcal{S} = \{s_1, s_2, \dots\}$, where an instance at time $t$ is either unlabelled $s_t = x_t \in \mathbb{R}^D$ or labelled $s_t = (x_t, y_t)$ with $y_t \in \{0, 1\}$. Unlabelled samples follow the marginal distribution $p_t(x)$, whereas labelled samples follow the joint distribution $p_t(x, y)$. Even though unlabelled examples are drawn from $p_t(x)$, we assume that they have a hidden, unknown label $y_t \in \{0, 1\}$ from $p_t(y|x)$. Concept drift, i.e., when $\exists t,\ p_{t+1}(x, y) \neq p_t(x, y)$, can cause data distributions to change over time affecting class priors, unconditional probability distribution, and/or conditional distributions [2]. Learners must be able to adapt to concept drift.

We assume learning systems that, at any time $t$, can store a sliding window $B_t = \{s_{t-N+1}, \dots, s_t\}$ of the latest $N$ samples to support learning. To enhance learning with few labels, Online Semi-Supervised Learning (OSSL) utilizes the advantage of structural presumptions that connect $p_t(x)$ and $p_t(y|x)$, such as cluster or manifold consistency, to update the learning model based on both labelled and unlabelled samples as the window slides by one.

Let the class prior at time $t$ be represented by $\pi_t(c) = p_t(y = c)$. Class imbalance happens when $\pi_t(\text{minority}) \ll \pi_t(\text{majority})$. As class priors could shift, the level of imbalance and the majority/minority roles may change over time. As unlabelled examples are assumed to belong to a given class, i.e., their input features $x_t$ are in regions of the problem space that belong to a given class, both supervised and unsupervised components may be impacted by imbalance in semi-supervised learning. For instance, error-driven updates are limited by the scarcity of labelled minority samples, and over time, biased decision boundaries may be reinforced by pseudo-labelling the unlabelled data as belonging to the majority class, encouraging further majority-class predictions. Under concept drift, where outdated representations further hinder minority recovery, class imbalance may become even more challenging.

## III. Related Work

### A. Online Class Imbalance Learning

Static cost-sensitive losses and global resampling are examples of traditional batch-oriented imbalance handling methods that rely on full supervision and stationary distributions [15]. These assumptions are rarely true in streaming environments. Global information pertaining the stream as a whole is unavailable, as examples arrive over time, while static cost-sensitive losses become unsuitable under changing imbalance ratios. Therefore, relying on these strategies in a streaming environment is either infeasible or could result in significant bias towards the majority class.

To overcome this issue, online class imbalance learning has been extensively researched in supervised environments. Online and chunk-based resampling, adaptive cost-sensitive learning, and ensemble-based streaming strategies are the three main types of current solutions [15], [16]. For instance, using local neighborhood data, streaming versions of SMOTE create synthetic minority samples [17]. Existing cost-sensitive learning approaches modify the decision rule to better handle imbalance [18]. Existing resampling-based ensemble [19]exhibit robustness under dynamic imbalance and explicitly model changing class proportions using time-decayed estimates. More recently, SMOClust proposed adaptive micro-clusters to oversample minority regions [20].

To estimate class distributions and find minority regions, these methods, however, depend on continuous access to ground-truth labels. This assumption does not hold in semi-supervised data streams with scarce labels. As a result, under limited supervision, these techniques cannot reliably identify or adjust to minority-class patterns.

### B. Online Semi-Supervised Learning under Concept Drift

Recent progress in OSSL illustrates the capability of utilizing extensive unlabelled data to adjust to changing data distributions. For example, by incorporating unlabelled samples into prototype updates and manifold regularization, Online Semi-supervised Neural Network (OSNN) [9] is a prototype-based neural model that updates the decision boundaries using labelled and unlabelled data through pseudo-labelling and manifold regularization, allowing adaptation to changes in $p(y \mid x)$ without the need of explicit drift detection. Temporal label propagation [7], micro-cluster-based frameworks [8], replay-based techniques for semi-supervised drifting stream learning with short lookback [10], and feature-evolving models for online semi-supervised learning with mix-typed streaming features [11] are examples of other drift-aware OSSL techniques based on different learning assumptions.

These techniques do not explicitly model class priors and assume that $p(y)$ remains stable over time, despite the fact that they successfully use unlabelled data to track evolving decision boundaries. As a result, under imbalance, majority-class dominance may gradually grow, particularly when minority samples are rarely labelled. Incorrect pseudo-labels in regions biased towards the majority-class can be used in model updates, further biasing the decision boundaries. This feedback process makes it difficult for the model to discover minority samples, especially when early minority labels are missing.

Addressing class imbalance in OSSL is challenging, as imbalance can directly influence the structural assumptions utilized by these approaches. For example, manifold and smoothness assumptions are based on the idea that nearby samples are likely to share the same label, but the shortage of labelled minority class examples can mean that their closest examples belong to the majority class. Clustering-based mechanisms are based on the assumption that class-

conditional regions are well-represented in the feature space. Under an imbalance scenario, samples from the majority class dominate local neighbourhood and density estimates. This may cause pseudo-label propagation and cluster formation leaning towards the majority class. As a result, simply integrating online class imbalance techniques such as those presented in Section III-A may not be enough. These techniques typically are developed for fully supervised settings and fail to take into account the error resulting from semi-supervised assumptions. Therefore, to address class imbalance in OSSL, new strategies that consider both supervised and unsupervised component concurrently may be necessary.

### C. Offline Class-Imbalanced Semi-Supervised Learning

Several semi-supervised learning techniques specifically address class imbalance in offline settings. ABC is an auxiliary balanced classifier trained with class-balanced sampling [21]. DARP recalibrates pseudo-label distributions to match class priors [22]. CReST gradually rebalances labelled sets using minority-biased pseudo-label selection [23]. Regularization is further altered by Suppressed Consistency Loss to reduce majority bias [24]. On static benchmarks, these methods show significant gains in balanced accuracy and minority recall.

However, none of these techniques can be directly applied to evolving data streams with unknown and drifting class priors, as they rely on the full training set being available beforehand, and static class distributions. In particular, they typically require access to an entire dataset to estimate global data manifolds or class-conditional feature distributions and to rebalance labelled and unlabelled subsets across training epochs. In streaming settings, where instances arrive over time and cannot be fully retained, these batch-level operations fail as they assume repeated passes over the full data and a stable global structure. Furthermore, when both class priors and decision boundaries change, it is difficult to keep reliable pseudo-label distributions and balanced representations without constantly re-estimating them, leading to significant memory and time overheads. This causes a challenge to directly adapt these methods to data streams.

### D. Research Gap

Overall, current online imbalance learning techniques assume complete supervision, whereas OSSL methods focus on adjusting to drift in $p(y \mid x)$ but overlook class imbalance and evolving class prior $p(y)$. Offline semi-supervised methods specifically deal with imbalance but cannot address streaming data. This highlights a gap between imbalance, semi-supervised adaptation, and online learning. The next sections of this paper reveal that this is a crucial gap, as class imbalance could affect semi-supervised learning even more severely than under fully labelled conditions.

## IV. Research Questions and Experimental Design

To better understand the significance of class imbalance in OSSL, we conduct the first study to characterize and quantify the sensitivity of OSSL to class imbalance compared to supervised data stream learning. The main research question being asked is: *How sensitive to class imbalance is OSSL compared to fully supervised online learning?* The study is based on OSNN [9] as a representative online semi-supervised neural learner, because it can learn from both labelled and unlabelled data online and uses techniques like pseudo-labelling and manifold regularization that are common in OSSL methods. We used 800 synthetic and 50 real-world binary data stream configurations capturing different drift speeds, severities, recurrences and underlying concepts.

### A. Data Streams

To answer our RQ, performance analysis must be conducted under carefully controlled experimental conditions that enable the isolation of interaction effects between important factors. We evaluate controlled combinations of labelling ratios (5%, 10%, 20%, 50%, and 100%), imbalance levels (minority proportions of 1%, 10%, 30%, and 50%), drift speeds (abrupt and gradual), and dataset families (SINE, SEA, STAGGER, and AGRAWAL). A 50% imbalance level is a balanced reference condition, and a 100% labelling is the fully supervised baseline that can be used to directly compare semi-supervised learning under the same imbalance regimes. Every stream is generated of a predetermined order of concepts capturing different drift severities, with 4,000 instances of each concept. For each concept sequence, two data stream versions are created: one based on a 400-instance transition window with instance-level interleaving representing gradual drifts, and one where each drift is abrupt, occurring in a single time step. In total, we adopt 10 concept sequences, 2 drift speeds (gradual and abrupt), 2 labelling strategies (uniform and cluster-based), 4 imbalance levels, and 5 labelling ratios yields, leading to $10 \times 2 \times 2 \times 4 \times 5 = 800$ synthetic data stream configurations. The code and data streams are available at https://github.com/haditalal/class-imbalance-online-semi-supervised-learning.

In accordance with standard OSSL protocols, labels are distributed throughout the stream. To guarantee temporal dispersion for sensitivity analysis, labelled samples are distributed uniformly through time in each concept segment. To evaluate the impact of the spatial coverage while maintaining the same label budget, we investigate both a uniform and a cluster-based space strategy. In the uniform strategy, labelled examples are sampled uniformly at random from the current concept. In the cluster-based strategy, for each class, KMeans ($k = 5$ or fewer if needed) divides the synthetic samples into clusters and each cluster gets label quota proportionally to its size. Then, the required number of labelled samples is chosen at random from each cluster to create the data stream.

In addition to synthetic benchmarks, we also adopt real-world streams from the USP Data Stream Repository [25]. Table I lists the chosen datasets and their main properties. The synthetic experiments enable us to fully control the concepts, whereas real-world datasets are employed to further check the identified trends on real world concepts and drifts. We use binary datasets directly, however we convert multi-class

TABLE I
REAL-WORLD DATA STREAMS FROM THE USP REPOSITORY [25].

| Dataset | Orig. | Minority(%) | #Feat. | #Samples | Drift Type |
|---|---|---|---|---|---|
| NOAA Weather | 2 | 31.38 | 8 | 18159 | Gradual, seasonal |
| Ozone | 2 | 6.32 | 72 | 2534 | Gradual, sensor-related |
| INSECTS-Incremental (Imb.) | 6 | 2.95 | 33 | 452044 | Incremental |
| INSECTS-Inc.-Abrupt-Recur. (Imb.) | 6 | 2.95 | 33 | 452044 | Abrupt + recurring |
| Keystroke | 4 | 24.95 | 10 | 1600 | Gradual, behavioral |

*Note:* Majority proportions are reported after one-vs-rest binarization for multi-class datasets.

datasets into binary streams using a one-vs-rest strategy by treating one class as the minority and all other classes as the majority. All data are processed in strict streaming order, and no future samples are used. To keep scales for distance calculations consistent, features are normalized online.

In the real-world evaluation, each run tests all datasets,labelling ratios, and both labelling strategies (uniform and cluster-based as described in Section IV-A adapted to continuous real-world streams without explicit concept boundaries) for OSNN. This enables us to study the effect of missing labels while maintaining the original real concepts intact. The total number of experimental configurations is obtained by combining 5 datasets, 2 labelling strategies, and 5 labelling ratios resulting in $5 \times 2 \times 5 = 50$ data stream configurations.

### B. Statistical Framework and Performance Metrics

we perform 30 runs per configuration with different random seeds, as this enables statistical analysis and allows the estimation of condition variability. A mixed (split-plot) ANOVA is applied over 30 independent runs for each synthetic data stream configuration, where the label ratio and imbalance level are the *within-subject factors*, and the dataset and drift speed are the *between-subject factors* (as they represent different relationships between inputs and outputs being present through the streams). This lets us analyze the main and interaction effects of imbalance, labelling ratio, and drift type on performance metrics statistically. In real-world data streams, imbalance levels and drift patterns are not directly managed to preserve realistic operational conditions. Therefore, a mixed ANOVA is performed with label ratio as the within-subject factor and dataset as the between-subject factor. In all of our analyses, Mauchly's tests of sphericity detected violations of the sphericity assumption (null hypothesis always rejected with p-value less than 0.001), so Greenhouse-Geisser corrections were used to guarantee conservative and statistically reliable conclusions. We focus our analysis on the geometric mean of recall and specificity (G-Mean), as it is not biased to the level of imbalance [26].

### C. Hyperparameter Tuning

Following existing work [9], for every experimental scenario defined by the dataset family, labelling ratio and imbalance level, hyperparameters are selected using random search over 1,000 configurations based on a separate tuning data stream constructed by mixing multiple streams from the same family under different drift types, using G-Mean as the selection criterion. To ensure a fair comparison, all scenarios employ the same set of randomly selected hyperparameter candidates. In the following evaluation phase, the chosen hyperparameter values are fixed for all data streams of the same family, labelling ratio and imbalance level. Testing occurs on data streams that are different from the tuning data stream, being generated with different random seeds across 30 repeated runs, ensuring that no data used for tuning are reused during performance evaluation.

For real-world experiments, we do not perform additional tuning, as no validation streams are available. Instead, we use a single hyperparameter setting selected from the synthetic study, namely the one that achieved the highest median G-Mean after aggregating 30 repeated runs for each of the 800 synthetic configurations. This choice avoids data leakage and ensures a fair evaluation, though it may not result in the best hyperparameter values for each individual real-world stream.

## V. ANALYSIS OF THE RESULTS

### A. Synthetic Data Stream Analysis

Table II shows the summary of the ANOVA results for both uniform and cluster-based labelling scenarios. As expected, the results demonstrate a significant main effect of Imbalance level on G-Mean ($p < .001$), with high effect size (partial $\eta^2 = 0.997$), validating that class imbalance is a strong determinant affecting model behaviour. The Label Ratio also has a strong main effect ($p < .001$, partial $\eta^2 = 0.981$), showing how important the presence of labelled data is. The main effect and the interactions between factors involving Drift Speed were not statistically significant or demonstrated minimal effect sizes.

The ANOVA results reveal a statistically significant three-way interaction among Imbalance, Label Ratio, and Dataset (uniform: $p < .001$, partial $\eta^2 = 0.883$; cluster-based: $p < .001$, partial $\eta^2 = 0.878$), demonstrating that the joint effect of imbalance and labelling is dataset-dependent and varies among various data-generating processes, with large effect size. Due to this significant interaction, results are plotted separately for each dataset family in Figure 1[1], for the uniform case. A table with the deltas in performance is in the supplementary material at https://github.com/haditalal/class-imbalance-online-semi-supervised-learning. The plots were very similar for the cluster-based case, and are thus omitted due to space constraints.

The plots show a clear difference between the supervised learning setting (Label = 100%) and the semi-supervised settings (Label $\leq$ 50%). Supervised learning always gets higher

[1]G-Means were averaged across data streams within each family because their G-Mean curves were similar.

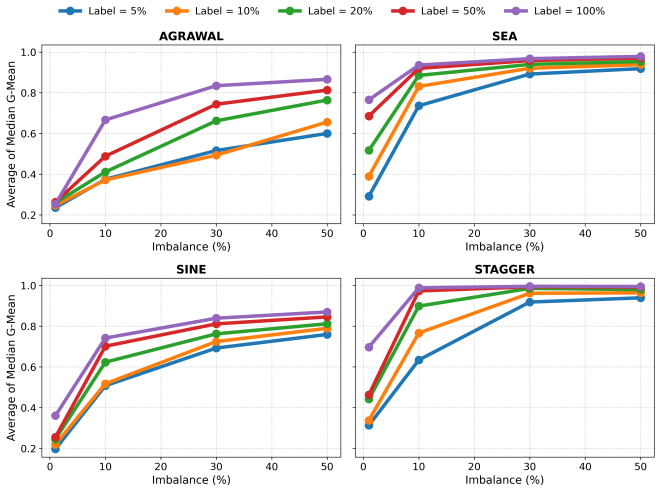


Fig. 1. Family-level interaction between imbalance level and labelling ratio under uniform labelling. For each dataset, curves show the mean of median G-Mean across datasets within each family over 30 runs.

G-Mean values and is much less sensitive to class imbalance. Even though performance gets worse as the minority class proportion decreases, this drop remains relatively small and only becomes more drastic once very small imbalance ratios (1%) are reached. Meantime, models trained on lower label ratios suffered drastic decreases in G-Mean earlier, at higher imbalance levels (especially SEA, STAGGER, and SINE to some extent) or presented a smooth decrease but with a much lower G-Mean starting point (AGRAWAL), despite using unlabelled data to support the learning process. Overall, depending on the dataset characteristics, class imbalance either results in consistently poor performance across labelling levels or amplifies the impact of limited labels.

These observations highlight a fundamental distinction between supervised and semi-supervised learning in the context of imbalanced data streams. Full supervision assists in fixing the imbalance by giving direct access to minority class labels, but semi-supervised learning can be potentially more vulnerable because it has access to less labelled minority class examples than in the fully supervised scenario, despite the class imbalance ratio itself being the same. The interaction between label scarcity and imbalance level shows patterns not addressed in previous studies solely concentrated on imbalanced supervised data stream learning or on balanced semi-supervised data stream learning. In particular, semi-supervised learning is unable to use the unlabelled data to sufficiently relieve the challenge posed by imbalance, possibly because the semi-supervised strategies are themselves biased in the imbalanced context.

### *B. Real-World Data Stream Analysis*

Table III shows the results of the ANOVA for OSNN with both uniform and cluster-based labelling for the real-world data. There is a significant main effect of Label Ratio in both labelling strategies ($p < .001$, partial $\eta^2 \approx 0.91$). This means that the amount of supervision available has a large effect on performance. This confirms that OSNN continues to be

TABLE II
G-MEAN: MIXED ANOVA SUMMARY UNDER UNIFORM AND CLUSTER-BASED LABELLING (GREENHOUSE–GEISSER CORRECTED FOR WITHIN-SUBJECT EFFECTS)

| | Uniform | | | Cluster-based | | |
|---|---|---|---|---|---|---|
| **Within-Subjects Effects** | | | | | | |
| **Effect** | **F** | $p$ | **Part.** $\eta^2$ | **F** | $p$ | **Part.** $\eta^2$ |
| Imbalance | 221703.794 | $<0.001$ | 0.997 | 221900.151 | $<0.001$ | 0.997 |
| Label Ratio | 29483.150 | $<0.001$ | 0.981 | 30616.128 | $<0.001$ | 0.981 |
| Imbalance × Dataset | 1000.311 | $<0.001$ | 0.939 | 832.059 | $<0.001$ | 0.928 |
| Imbalance × Label Ratio × Dataset | 485.428 | $<0.001$ | 0.883 | 464.671 | $<0.001$ | 0.878 |
| Label Ratio × Dataset | 124.179 | $<0.001$ | 0.658 | 204.435 | $<0.001$ | 0.760 |
| Imbalance × Label Ratio | 529.116 | $<0.001$ | 0.477 | 555.185 | $<0.001$ | 0.489 |
| Imbalance × Dataset × Drift Speed | 1.864 | 0.032 | 0.028 | 0.338 | 0.978 | 0.005 |
| Label Ratio × Dataset × Drift Speed | 0.936 | 0.572 | 0.014 | 0.472 | 0.995 | 0.007 |
| Imbalance × Label Ratio × Dataset × Drift Speed | 1.033 | 0.413 | 0.016 | 0.472 | 0.998 | 0.007 |
| Imbalance × Label Ratio × Drift Speed | 0.898 | 0.481 | 0.002 | 0.189 | 0.960 | 0.000 |
| Label Ratio × Drift Speed | 0.416 | 0.779 | 0.001 | 0.290 | 0.869 | 0.000 |
| Imbalance × Drift Speed | 0.048 | 0.902 | 0.000 | 0.129 | 0.775 | 0.000 |
| **Between-Subjects Effects** | | | | | | |
| Dataset | 32978.104 | $<0.001$ | 0.998 | 36762.128 | $<0.001$ | 0.998 |
| Dataset × Drift Speed | 1.362 | 0.202 | 0.021 | 0.242 | 0.988 | 0.004 |
| Drift Speed | 0.030 | 0.862 | 0.000 | 0.191 | 0.662 | 0.000 |

TABLE III
G-MEAN: MIXED ANOVA SUMMARY FOR OSNN ON REAL-WORLD STREAMS (GREENHOUSE–GEISSER CORRECTED)

| | Uniform | | | Cluster-based | | |
|---|---|---|---|---|---|---|
| **Within-Subjects Effects** | | | | | | |
| **Effect** | **F** | $p$ | **Part.** $\eta^2$ | **F** | $p$ | **Part.** $\eta^2$ |
| Label Ratio | 1450.313 | $<0.001$ | 0.909 | 1536.753 | $<0.001$ | 0.914 |
| Label Ratio × Dataset | 291.332 | $<0.001$ | 0.889 | 170.896 | $<0.001$ | 0.825 |
| **Between-Subjects Effects** | | | | | | |
| Dataset | 24077.914 | $<0.001$ | 0.998 | 13270.888 | $<0.001$ | 0.997 |

influenced by a lack of labels in real-world streams as it is in synthetic data. The interaction between Label Ratio and Dataset is significant for both labelling strategies ($p < .001$, partial $\eta^2 = 0.889$ for uniform and $0.825$ for Cluster-based), indicating that the effect of label scarcity varies depending on the dataset. This means that the amount of degradation that OSNN experiences in semi-supervised settings differs significantly between streams with different feature distributions and drift characteristics. The notable between-subject effect of Dataset further confirms that baseline difficulty differs significantly among the chosen real-world streams, corroborating the results obtained with the synthetic datasets.

However, the G-Mean does not increase consistently with increasing label ratio when uniform labelling is employed (Figure 2). Instead, several datasets show fluctuations and occasional drops. This may be due to noise or local variations in class distributions. Compared to the cluster-based scenario, uniform labelling was more vulnerable in complex real-world streams, possibly because the imbalance ratios are not constant across local regions or clusters, which may amplify this effect. Deeper analysis is needed to confirm the reasons underlying these observations, but these results suggest that new strategies proposed to tackle class imbalance in semi-supervised learning may need to consider the distribution of labels across space, which is particularly challenging in streaming scenarios where

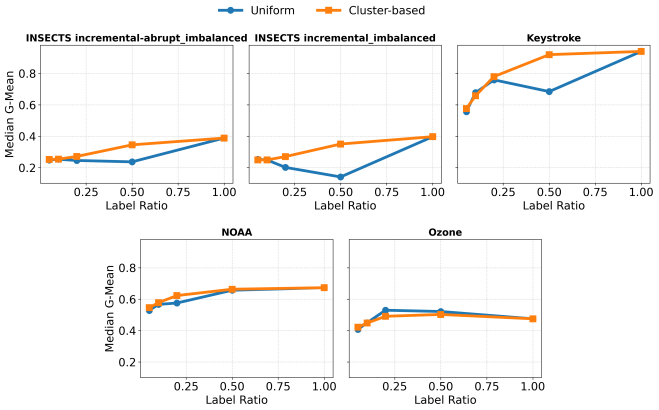


Fig. 2. Comparison between uniform and cluster-based labelling strategies for OSNN on real-world data streams. Curves show median G-Mean over 30 runs for each dataset across different label ratios.

the full training set is not available beforehand.

## VI. Conclusions

Despite the prevalence of both limited labelling and class imbalance in many real world data stream applications, our study showed that performance of OSSL can be severely impacted by class imbalance. This effect was found to be more significant in semi-supervised environments compared to fully supervised online learning, despite the imbalance ratio being the same. Semi-supervised data stream learning strategies themselves may be affected by class imbalance, such that their ability to cope with class imbalanced data streams can be severely affected. Moreover, the labelling distribution plays a significant impact on the performance in semi-supervised class imbalanced data streams, which poses a particular challenge in streaming environments where the whole data set is not available beforehand to estimate such distribution.

The proposal of novel semi-supervised strategies able to cope with class imbalance in data streaming environments is thus a significant future direction of study. New strategies that can benefit from semi-supervised assumptions such as manifold, clustering and smoothness in imbalanced streaming scenarios are necessary. In particular, future work should look into novel strategies to benefit from unlabelled data to *cope with* class imbalance, rather than propagating the majority class bias through them. Our study could also be extended to investigate other learning approaches, multi-class problems and problems with verification latency (label delay).

## References


[1] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, "Learning in non-stationary environments: A survey," *IEEE CIM*, vol. 10, no. 4, pp. 12–25, 2015.

[2] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM CSUR*, vol. 46, no. 4, pp. 1–37, 2014.

[3] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection and concept-drift adaptation with delayed supervised information," in *2015 international joint conference on Neural networks (IJCNN)*. IEEE, 2015, pp. 1–8.

[4] Z. Ding, S. Qiu, Y. Guo, J. Lin, L. Sun, D. Fu, Z. Yang, C. Li, Y. Yu, L. Meng *et al.*, "Labelecg: A web-based tool for distributed electrocardiogram annotation," in *International Workshop on Machine Learning and Medical Engineering for Cardiovascular Healthcare*. Springer, 2019, pp. 104–111.

[5] T. Braun, I. Pekaric, and G. Apruzzese, "Understanding the process of data labeling in cybersecurity," in *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 2024, pp. 1596–1605.

[6] K. B. Dyer, R. Capo, and R. Polikar, "Compose: A semisupervised learning framework for initially labeled nonstationary streaming data," *IEEE TNNLS*, vol. 25, no. 1, pp. 12–26, 2013.

[7] T. Wagner, S. Guha, S. Kasiviswanathan, and N. Mishra, "Semi-supervised learning on data streams via temporal label propagation," in *ICML*, 2018, pp. 5095–5104.

[8] S. U. Din, J. Shao, J. Kumar, W. Ali, J. Liu, and Y. Ye, "Online reliable semi-supervised learning on evolving data streams," *Information Sciences*, vol. 525, pp. 153–171, 2020.

[9] R. G. F. Soares and L. L. Minku, "Osnn: An online semisupervised neural network for nonstationary data streams," *IEEE TNNLS*, vol. 34, no. 9, pp. 6029–6041, 2023.

[10] W. Ren, P. Wang, X. Li, C. E. Hughes, and Y. Fu, "Semi-supervised drifted stream learning with short lookback," in *KDD*, 2022, pp. 1504–1513.

[11] D. Wu, S. Zhuo, Y. Wang, Z. Chen, and Y. He, "Online semi-supervised learning with mix-typed streaming features," in *AAAI*, vol. 37, no. 4, 2023, pp. 4720–4728.

[12] Y. Guo, J. Pu, B. Jiao, Y. Peng, D. Wang, and S. Yang, "Online semi-supervised active learning ensemble classification for evolving imbalanced data streams," *Applied Soft Computing*, vol. 155, p. 111452, 2024.

[13] European Banking Authority and European Central Bank, "Report on payment fraud," European Banking Authority and European Central Bank, Tech. Rep., 2024. [Online]. Available: https://www.eba.europa.eu/sites/default/files/2024-08/465e3044-4773-4e9d-8ca8-b1cd031295fc/EBA_ECB%202024%20Report%20on%20Payment%20Fraud.pdf

[14] Y. K. Kim, W.-D. Seo, S. J. Lee, J. H. Koo, G. C. Kim, H. S. Song, and M. Lee, "Early prediction of cardiac arrest in the intensive care unit using explainable machine learning: retrospective study," *J. Med. Internet Res.*, vol. 26, p. e62890, 2024.

[15] S. Wang, L. L. Minku, and X. Yao, "A systematic study of online class imbalance learning with concept drift," *IEEE TNNLS*, vol. 29, no. 10, pp. 4802–4821, 2018.

[16] G. Aguiar, B. Krawczyk, and A. Cano, "A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework," *Machine Learning*, vol. 113, no. 7, pp. 4165–4243, 2024.

[17] Ł. Korycki and B. Krawczyk, "Online oversampling for sparsely labeled imbalanced and non-stationary data streams," in *IJCNN*, 2020, pp. 1–8.

[18] L. Loezer, F. Enembreck, J. P. Barddal, and A. de Souza Britto Jr, "Cost-sensitive learning for imbalanced data streams," in *SAC*, 2020, pp. 498–504.

[19] S. Wang, L. L. Minku, and X. Yao, "Resampling-based ensemble methods for online class imbalance learning," *IEEE TKDE*, vol. 27, no. 5, pp. 1356–1368, 2014.

[20] C. W. Chiu and L. L. Minku, "Smoclust: synthetic minority oversampling based on stream clustering for evolving data streams," *Machine Learning*, vol. 113, no. 7, pp. 4671–4721, 2024.

[21] H. Lee, S. Shin, and H. Kim, "Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning," in *NeurIPS*, vol. 34, 2021, pp. 7082–7094.

[22] J. Kim, Y. Hur, S. Park, E. Yang, S. J. Hwang, and J. Shin, "Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning," *NeurIPS*, vol. 33, pp. 14 567–14 579, 2020.

[23] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang, "Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning," in *CVPR*, 2021, pp. 10 857–10 866.

[24] M. Hyun, J. Jeong, and N. Kwak, "Class-imbalanced semi-supervised learning," *arXiv*, 2020.

[25] V. M. A. Souza, D. M. Reis, A. G. Maletzke, and G. E. A. P. A. Batista, "Challenges in benchmarking stream learning algorithms with real-world data," *Data Mining and Knowledge Discovery*, vol. 34, pp. 1805–1858, 2020.

[26] A. Luque, A. Carrasco, A. Martin, and A. De las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019.