Evaluating supervised concept drift detection for online imputation on wearable data streams with changing missing mechanisms

Afonso M. S. Lima[1], **Matheus Q. Mota**[1], **Leonardo S. Cardoso**[1], **Elaine P. M. de Sousa**[1]

[1]Institute of Mathematical and Computer Science (ICMC)
University of São Paulo (USP)
São Carlos – SP – Brazil

{afonso.matheus, matheusqueirozmota, leo.silva.cardoso}@usp.br
parros@icmc.usp.br

Abstract. *Wearable data streams often contain missing values, for which there is uncertainty about the cause of the missingness. Further, this missingness may also change throughout the stream processing. Our study evaluates the effectiveness of concept drift detectors applied in a supervised manner for online imputation on heart rate data streams collected from wearable devices. Experiments were conducted under changing missing mechanisms scenarios, utilizing the Hoeffding Adaptive Tree regressor alongside three concept drift detectors for monitoring model error: Page-Hinkley, Kolmogorov-Smirnov Windowing, and Adaptive Windowing. We also compared it with baselines Hoeffding Tree and mean imputation. Our results show that supervised concept drift detection yields worse performance than non-aware concept drift approaches. This indicates that supervised detectors may be insufficient to address data distribution changes caused by missing values, highlighting the need for further research.*

1. Introduction

The use of wearable devices has increased drastically over the years, generating high amounts of streaming data. However, continuously generated wearable data streams often contain missing values, which can lead to biased findings and incorrect treatment decisions [Isgut et al. 2022]. This challenge is compounded by uncertainty regarding the cause of the missing values, i.e., the missing mechanism, as missingness may arise from random failures, dependence on other observed features, or reflect characteristics absent from available data [Ren et al. 2023, Benhamza et al. 2025]. Furthermore, missingness assumptions are challenging to handle due to the non-stationary nature of data streams. Online Machine Learning (OML) encompasses incremental or online learning methods specifically designed to efficiently process streaming data while being robust to concept drift, i.e., changes in incoming data patterns [Bifet et al. 2023, Bartz and Bartz-Beielstein 2024]. Concept drift detection techniques mostly operate in a supervised manner, detecting changes based on the model's errors [Mahdi et al. 2024, Lukats et al. 2025].

Acknowledgments This research was in part funded by the Brazilian Coordination for Improvement of Higher Level Personnel (CAPES) - grant 001, the National Research Council (CNPq), and the São Paulo Research Foundation (FAPESP) - grants 2023/18026-8, 2024/13328-9, 2025/27722-3, and 2026/12389-0.

This study evaluates supervised concept drift detection for online imputation on wearable data streams across three simulated scenarios with changing missing mechanisms. The simulated scenarios reflect changing patterns similar to sudden, recurrent, and gradual drifts, but applied to the missing values distribution, following the three main missing mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR).

We experimented with heart rate data from 30 patients (i.e., 30 data stream sets) [Mishra et al. 2020]. Three concept drift detectors were evaluated in combination with the Hoeffding Adaptive Tree (HAT) regressor for online imputation, alongside a baseline mean imputation and Hoeffding Tree (HT) non-concept drift aware regressor, across three simulated missingness scenarios with varying mechanisms. Imputation performance was evaluated for each scenario. Our results show that supervised concept drift detection yields worse performance than non-aware concept drift approaches.

2. Background

The increasing use of wearable devices, the Internet-of-Things technologies and mobile health applications raises the issue of real-time streaming data imputation [Lima and Sousa 2024, Benhamza et al. 2025]. Missing values may arise from random sensor faults or inconsistent collection periods, such as varying compliance and wearing behavior, corresponding to distinct missing mechanisms that require different imputation solutions [Mishra et al. 2020]. Table 1 summarizes each mechanism definition along with a practical healthcare example. Such missing information may degrade data analysis performance and yield biased results [Benhamza et al. 2025].

Table 1. Missing mechanisms summary and healthcare example [Lima 2025]

Missing Mechanism	Definition	Example
Missing Completely At Random (MCAR)	The missing cause is unrelated to other observed or unobserved values.	The wearable device failed, and the value was not generated.
Missing At Random (MAR)	The missing cause depends on other observed attributes.	Patient removed their wearable sensor in sleep hours (“heartrate” depends on “datetime”).
Missing Not at Random (MNAR)	The missing cause is related to specific information about the own missing value that is not present in the dataset.	The wearable sensor is not sensitive enough to low values, setting it to zero. (“heartrate” depends on itself).

Futhermore, as batch approaches may be insufficient to meet stream processing requirements, OML solutions apply incremental learning to fit machine learning models into these wearable data streams [Bifet et al. 2023, Bartz and Bartz-Beielstein 2024]. Generally, those methods are designed considering that: 1) each instance can be used only once; 2) processing time is severely limited; 3) memory is limited; 4) the algorithm must be able to return a result at any time; 5) data streams are assumed to change properties over time, such as distribution, recurrence, velocity, and time interval, and this phenomenon is known as concept drift [Bartz and Bartz-Beielstein 2024, Mahdi et al. 2024]. Therefore, OML methods are more effective at processing new incoming data, detecting changes in real time, and eliminating the need to retrain the entire model. Most concept drift detectors proposed in the literature operate in a supervised manner, as they monitor the predictive performance of a classifier deployed on the data stream. [Lukats et al. 2025].

The non-stationary nature of data streams may extend to the underlying missing distribution, resulting in diverse mechanism settings throughout stream processing. This observation raises the question of how current OML methods address potential changes in the missing mechanism settings. When considering only one missing mechanism, there is evidence that concept drift-aware methods achieve better, more stable results for imputation and mining task improvement [Lima 2025, Lima 2026]. To the best of our knowledge, no prior study has investigated changing missing mechanism scenarios in online data stream processing tasks, highlighting the need for initial experimentation in this area, given data streams' evolving characteristics.

3. Method

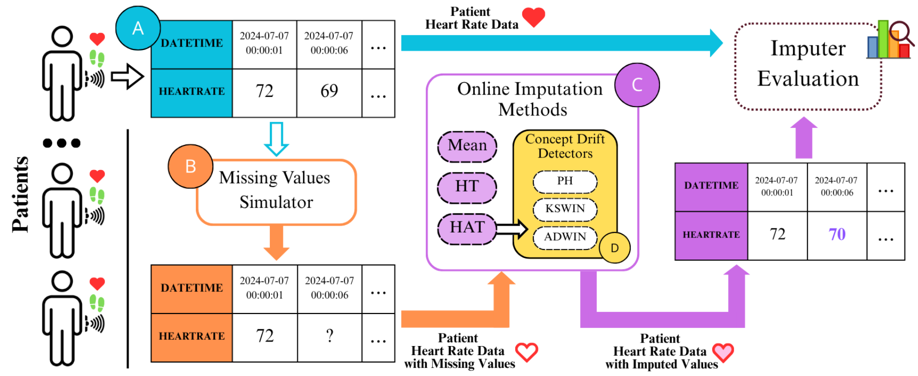


Figure 1. Method for online imputation analysis with concept drift detectors

Figure 1 provides an overview of our method. We based our experiments on the dataset presented by Mishra *et al.* [Mishra et al. 2020]. The patient's wearable data (Fig. 1-A) consists of several days of heart rate data for 30 patients, with measurement counts ranging from thousands to millions for each patient, sampled at short intervals (e.g., 5 seconds). Each measurement has a timestamp (e.g., "2024-07-07 00:00:01") and the respective heart rate value (e.g., "72", "69"). We simulate missing values following scenarios that apply two or more of the main missing mechanisms (Fig. 1-B). Across all scenarios, missingness was simulated using the *mdatagen* Python library [Santos et al. 2019, Mangussi et al. 2024] as follows: for MCAR, the "random" strategy was used, in which missing values are randomly selected; for MAR, we simulate a dependency on the hour in the "datetime" feature, with the early hours of the day likely to have missing values ("lowest" strategy); for MNAR, we simulate dependence on the highest heart rate values (threshold = 1.0). These parameter settings are based on previous studies that simulated scenarios in which missing values arise for different reasons, such as random sensor faults and inconsistent collection periods (e.g., varying compliance and patient wearing behavior) [Lima 2025, Lima 2026].

Aiming to simulate different situations based on types of concept drift [Mahdi et al. 2024, Lukats et al. 2025], we defined three missingness scenarios: Sce_1 **- abrupt changes in the missing distribution**) The MAR setting is simulated for the first third of each patient's heart rate set, followed by MCAR for the second third, and MNAR for the final third, each part with a 10% missing rate. $\boldsymbol{Sce_2}$ **- gradual change in the missing distribution**). The MAR setting is applied in the first quarter, then MNAR in the second, MAR in the third, and MNAR in the final quarter, with each part having a 7.5% missing rate. $\boldsymbol{Sce_3}$ **- recurrent changes in the missing distribution**) MNAR is applied to the first fifth, followed by MAR in the second fifth, and then alternating MNAR, MAR, and MNAR, each with a 6% missing rate. We defined a total missing rate of 30% across all scenarios, ensuring that each mechanism partition has both an equal missing ratio and a minimum number of missing occurrences. Figure 2 exemplifies each scenario on the dataset of patient labeled as "AAXAA7Z". The online imputation models are built upon the available heart rate data (i.e., the gray dots) and tested on the simulated missing heart rate data (i.e., colored dots).

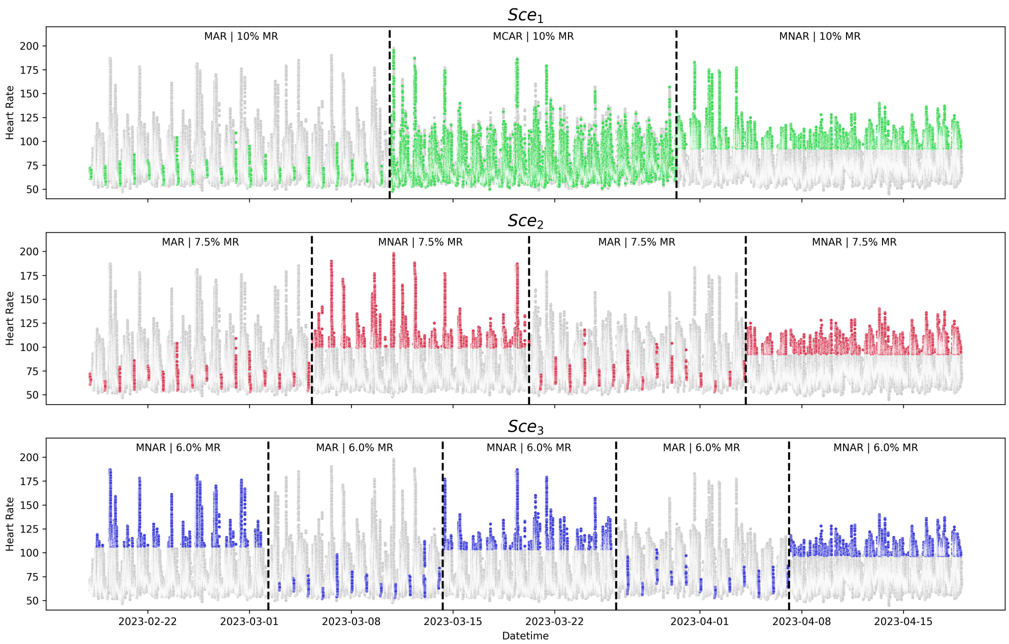


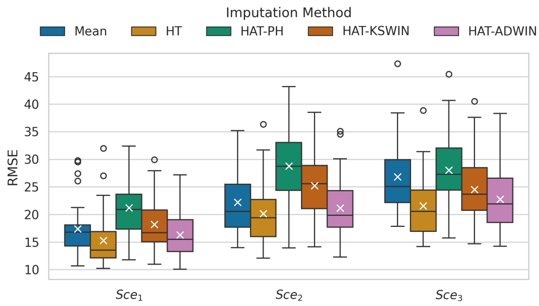
Figure 2. Missingness scenarios applied to patient "AAXAA7Z" wearable data. Colored dots indicate heart rate measurements that the mechanism setting designates as missing. Vertical dotted lines illustrate the moment of mechanism change.

Online imputation was performed using methods implemented in River, a Python library designed for machine learning on streaming data [Montiel et al. 2021, Bartz and Bartz-Beielstein 2024]. We employed the HAT regressor, which supports integration with various concept-drift detectors. For comparison purposes, a HT regressor and a baseline mean imputation were also analyzed (Fig.1-C). We evaluated three concept drift detectors in conjunction with HAT: Page-Hinkley (PH), Kolmogorov-Smirnov Windowing (KSWIN), and Adaptive Windowing (ADWIN) [Montiel et al. 2021, Lukats et al. 2025] (Fig.1-D). These concept drift detectors work alongside HAT in a supervised manner: they detect changes based on the model's errors, signaling that an update to the current model is due [Mahdi et al. 2024]. In this imputation task, we incrementally test-then-train models on each available heart rate measurement, while the concept drift detectors monitor the models' errors. When a missing value occurs (i.e, "heartrate" is NaN), the current imputation model is apllied.

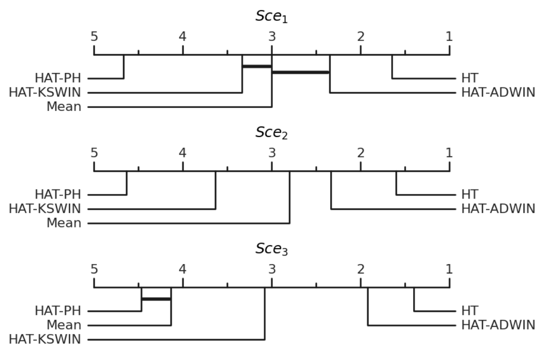
Finally, for each identified missing value, we calculate the Root Mean Squared Error (RMSE) between the original and the imputed value for each patient. Parameter tuning for the concept drift detectors was performed for each missingness scenario using patient subsets (i.e., 1, 3, and 5 randomly selected patients), stopping when the parameter converged and due to resource limitations. To ensure that sequential mechanisms' partitions have distributions that differ significantly, the Mann–Whitney U test and Levene's test were applied to compare distributions after missing values simulation. For all patients and missingness scenarios, at least one statistical test rejected the null hypothesis across all sequential partitions. Parameter grids, p-value tables, and codes for reproducibility are available in a public repository[1].

4. Results

Figure 3 aggregates the experimental results. Figure 3a presents boxplots with the imputer's mean RMSE for each of the 30 patients in each missingness scenario. Figure 3b presents the critical difference diagram for ranking imputation methods. This provides a visualization of method rankings and determines whether observed performance differences are statistically significant by applying the Wilcoxon test to paired RMSE results, with Holm's correction to control Type I error [Ismail Fawaz et al. 2019].



(a) Boxplots with mean RMSE results for each of the 30 patients for each imputation method (Mean, HT, HAT-PH, HAT-KSWIN, HAT-ADWIN) in each missingness scenario. The white X indicates the mean.



(b) Critical difference diagrams for ranking imputation methods using the mean RMSE from each patient. The rightmost methods achieved higher imputation rankings. The bold line marks methods not statistically different.

Figure 3. Experimental results.

The results show that concept drift solutions performed worse than the HT baseline, which ranked first across all missingness scenarios. The HT obtained not only the lowest RMSE results but also the most stable results. These findings suggest that the identified concept drifts, which indicate when to construct a new model in HAT, do not yield improved performance over the HT approach, which adapts models solely to the most recent data. Therefore, detecting concept drift through model errors was insufficient to enhance model adaptation. Compared to the mean baseline, the concept drift aware methods range from equivalent to better performance. From worst to best, the concept drift detectors applied with HAT rank as: HAT-PH, HAT-KSWIN, and HAT-ADWIN.

[1] https://github.com/afonsoMatheus/supervised-concept-drift-analysis

HAT-PH detects concept drift by monitoring deviations in the models' accumulated mean error, which may generate both false-positive and false-negative signals for model updates due to high variance in the distribution. HAT-KSWIN detects changes in model errors using a non-parametric statistical test on two sub-windows: one smaller, containing the most recent objects, and the other larger, containing older objects. Because older objects are uniformly selected for comparison with newer objects, concept drift detection may suffer on highly variable data. HAT-ADWIN, which automatically resizes the window to better observe changes, obtained better results because it can respond more effectively to current changes in the model's error. Additionally, ADWIN parameter tuning is simpler than tuning both PH and KSWIN, which have more threshold and window-size variables.

Furthermore, when missingness scenarios entail more mechanism changes (i.e., Sce_2 and Sce_3), the ADWIN detector obtains better results, as more model updates may be necessary to adapt to the current data state. Looking at the mean baseline, multiple mechanisms changes with highly different missing distributions reduced its effectiveness in Sce_2 and Sce_3, whereas both HT and HAT performed better due to the model update capacity. Futhermore, parameter tuning plays a critical role in the effectiveness of concept drift. We employed parameter tuning for each missingness scenario, but patient-based tuning could greatly improve results. However, this approach is infeasible for real-time applications, further motivating the use of ADWIN on new proposals due to its adaptive capabilities and less cryptic instantiation.

5. Conclusion

This study evaluated three concept drift detectors applied in a supervised manner for online imputation on wearable data streams across three simulated missingness scenarios with changing missing mechanisms.

Our findings indicate that concept drift detection, when applied to detect errors in model performance, may be insufficient to address data distribution changes caused by missing values, given that it yielded worse performance than non-aware concept drift approaches in our experimental settings. Further, as more missing mechanism changes occur in each missingness scenario, concept drift detection methods demonstrate improved performance compared to baseline mean imputation. Notably, the ADWIN concept drift solution showed the most promising results in terms of performance and more straightforward parameter tuning due to adaptive window size. However, these methods continue to underperform compared to a non-concept drift aware online regressor, which continuously adapts the same model. This indicates that the concept drifts detected by the supervised approach did not improve imputation by rebuilding the model at each drift.

As future work, further experimentation is needed to assess models' robustness to diverse missingness scenarios in real-time problems, varying missing rates and mechanisms, in addition to evaluating model performance across processing for individual patient analysis. Evaluating the impact of imputation on improving the online mining task is also relevant, as even better imputation models may not be sufficient to mitigate the effect of missing values, motivating the conception of missingness aware solutions.

Regarding generative AI usage. Generative AI tools were used solely for English language editing (grammar and style). All content was reviewed, revised, and approved by the authors, who assume full responsibility for the manuscript.

References

Bartz, E. and Bartz-Beielstein, T. (2024). *Online Machine Learning*. Springer.

Benhamza, K., Benselim, R., Naidja, H., and Seridi, H. (2025). A comprehensive survey of imputation methods in medical missing data analysis. *Applied Intelligence*, 55(11):778.

Bifet, A., Gavalda, R., Holmes, G., and Pfahringer, B. (2023). *Machine learning for data streams: with practical examples in MOA*. MIT press.

Isgut, M., Gloster, L., Choi, K., Venugopalan, J., and Wang, M. D. (2022). Systematic review of advanced ai methods for improving healthcare data quality in post covid-19 era. *IEEE Reviews in Biomedical Engineering*, 16:53–69.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963.

Lima, A. M. S. (2026). The impact of missing values on anomaly detection tasks in healthcare wearable data. *Journal of Information and Data Management*. Accepted for publication; to appear.

Lima, A. S. (2025). Assessing the impact of missing value mechanisms on anomaly detection in healthcare wearable data. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 781–787, Brasil.

Lima, A. S. and Sousa, E. (2024). Handling missing values in data streams: An overview. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 750–756, Brasil.

Lukats, D., Zielinski, O., Hahn, A., and Stahl, F. (2025). A benchmark and survey of fully unsupervised concept drift detectors on real-world data streams. *International Journal of Data Science and Analytics*, 19(1):1–31.

Mahdi, O. A., Ali, N., Pardede, E., Alazab, A., Al-Quraishi, T., and Das, B. (2024). Roadmap of concept drift adaptation in data stream mining, years later. *IEEE Access*, 12.

Mangussi, A. D., Santos, M. S., Lopes, F. L., Pereira, R. C., Lorena, A. C., and Abreu, P. H. (2024). mdatagen: A python library for generating missing data. https://arthurmangussi.github.io/pymdatagen/.

Mishra, T., Wang, M., Metwally, A. A., Bogu, G. K., Brooks, A. W., Bahmani, A., Alavi, A., Celli, A., Higgs, E., Dagan-Rosenfeld, O., et al. (2020). Pre-symptomatic detection of covid-19 from smartwatch data. *Nature biomedical engineering*, 4(12):1208–1220.

Montiel, J., Halford, M., Mastelini, S. M., Bolmier, G., Sourty, R., Vaysse, R., Zouitine, A., Gomes, H. M., Read, J., Abdessalem, T., et al. (2021). River: machine learning for streaming data in python. *Journal of Machine Learning Research*, 22(110):1–8.

Ren, L., Wang, T., Seklouli, A. S., Zhang, H., and Bouras, A. (2023). A review on missing values for main challenges and methods. *Information Systems*, page 102268.

Santos, M. S., Pereira, R. C., Costa, A. F., Soares, J. P., Santos, J., and Abreu, P. H. (2019). Generating synthetic missing data: A review by missing mechanism. *IEEE Access*, 7:11651–11667.