Handling of missing values in wearable data streams

Afonso M. S. Lima[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@usp.br parros@icmc.usp.br

Level: PhD's in Computer Science and Computational Mathematics
Admission: 07/2022 **Qualifying Exam:** 03/04/2024 **Defense:** 03/2027
Completed Activities: Completion of mandatory credits, qualifying exam and presentation, foreign language proficiency exam, data collection, data exploratory analysis, missing values impact assessment
Future Activities: Missing values handler method conception, method implementation, experimental tests with correlated approaches, results evaluation and discussion
Publications: Handling missing values in data streams: An overview. [Lima and Sousa 2024]

Abstract. *The increasing volume of data generated by interconnected devices has amplified the need for efficient stream mining methods, particularly in healthcare applications using wearable devices. These systems enable continuous health monitoring and support early interventions. However, missing values—common in streaming data—can lead to biased or invalid decisions, especially when their underlying causes are unknown. This PhD project proposes a preprocessing method to handle missing values in healthcare wearable data streams, addressing challenges such as data evolution, diverse missing mechanisms, and computational constraints. The goal is to improve data quality and the robustness of mining tasks in real-time health monitoring systems.*

Resumo. *O aumento do volume de dados gerados por dispositivos interconectados intensificou a necessidade por métodos eficientes de mineração de fluxo de dados, especialmente em aplicações de saúde com uso de dispositivos vestíveis (e.g., smartwatchs). Esses sistemas permitem o monitoramento contínuo da saúde e apoiam diagnósticos mais rápidos. No entanto, a presença de valores ausentes, comum em fluxo de dados, pode levar a decisões enviesadas ou inválidas, especialmente quando suas causas são desconhecidas. Este projeto de doutorado propõe um método de pré-processamento para tratar valores ausentes em fluxos de dados provenientes de dispositivos vestíveis na área da saúde, abordando desafios como a evolução dos dados, diferentes mecanismos de ausência e restrições computacionais. O objetivo é melhorar a qualidade dos dados e a robustez das tarefas de mineração em sistemas de monitoramento em tempo real.*

Acknowledgments This research was supported by CAPES (Brazilian Coordination for Improvement of Higher Level Personnel) and CNPq (Brazilian National Council for Supporting Research).

1. Introduction

Remote health monitoring using wearable devices, such as smartwatches, holds great promise in healthcare by allowing for continuous monitoring of individuals, facilitating early detection, preventive care, and timely interventions [Getzen et al. 2023]. However, this approach is significantly affected by missing values, as it is common that continuously generated wearable data presents a percentage of missingness [Psychogyios et al. 2023], which can yield biased findings and mistaken treatment decisions [Isgut et al. 2022].

However, this and other application domains are significantly affected by missing values, as it is common that this continuously generated data presents a percentage of missingness [Psychogyios et al. 2023]. As patient-generated health data obtained from wearable devices can guide clinicians in decision-making for optimal patient treatment, missing data in this context can yield biased findings and unfair treatment decisions, leading to failure in decision-making [Isgut et al. 2022].

In this context, this PhD project aims to develop a preprocessing method to handle missing values in healthcare wearable data streams, focusing on mining tasks. We consider the gaps identified in the state of the art, such as the ability to handle different types of data evolution and different missing value mechanisms, in addition to meeting computational processing and memory requirements necessary for the data stream context.

Therefore, the results of this research project may enhance the management of missing data within the healthcare sector, thereby making data stream mining tasks more resilient to the presence of missing data. Our research can contribute to computer science and medical fields, advance stream mining methods, improve database quality, and be applied to healthcare. We aim to operate effectively in theoretical and practical dimensions.

1.1. Problem Description

There are two main algorithmic challenges when dealing with streaming data: fast large data generation and real-time processing requirements. Any step of the knowledge discovery process (e.g., preprocessing, validation, etc.) has to consider these requisites when conceiving a new method to process data streams efficiently [Lima and Sousa 2024]. Additionally, due to the dynamic nature of data streams, any learning model should adapt to evolving data, known as concept drift [Bahri et al. 2021].

Most research in data streams does not address the problem of missing values, imputing missing values in data streams remains a relatively unexplored issue. Table 1

Table 1. Accessible research related to missing values handling aspects

Reference	Data Stream Requisites	Concept Drift Exploration	Missing Mechanism
Fountas and Kolomvatsos (2020)	Not considered	Not explored	Not defined
Sun et al. (2020)	Not considered	Not explored	Not defined
Dong et al. (2021)	Considered	Explored	Not defined
Zhang and Thorburn (2022)	Not considered	Not explored	MAR
Halder et al. (2022)	Considered	Poorly explored	Not defined
Liu et al. (2023)	Considered	Not explored	MCAR, MAR
Li et al. (2023)	Considered	Not explored	Not defined

summarizes the accessible research and missing values handling aspects for data streams [Lima and Sousa 2024]. This scenario of few studies with diverse approaches indicates that missing value imputation in data streams is a new and currently open issue in the stream mining field. Furthermore, current solutions fail to fully cope with essential aspects of handling missing values in data streams, namely: data stream requisites, concept drift exploration and the missing mechanism assumption [Lima and Sousa 2024].

Table 2. Missing mechanisms summary and healthcare example

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).

Further, it is challenging to understand the cause of the missing values on the missing mechanism assumption, as they can be complex and influenced by external factors. For instance, in healthcare wearable data, in addition to random sensor faults, these missing values may occur due to inconsistent data collection periods (e.g., wearing behavior and compliance vary by person) [Mishra et al. 2020]. Table 2 presents the three principal missing mechanisms and a practical example in the healthcare domain. Correctly identifying the missing mechanism is crucial for better treatment of missing values [Ren et al. 2023].

1.2. Proposed solution

The main objective of this PhD project is to conceive and implement a method to handle missing values in data streams, considering processing requisites, concept drift exploration, and missing mechanism assumptions. This new method will be experimented with data from the healthcare domain. To support this goal, we have the following specific objectives: 1) Assess how missing values affect the efficiency of data stream mining tasks; 2) Evaluate various data delimiting techniques to represent and process a data stream; 3) Analyze the performance of imputation methods based on different assumptions about the mechanisms of missing values. To achieve these, the following hypotheses are formulated:

1. Considering the correlation between attributes, if present, when imputing missing values in multidimensional data streams increases the accuracy of the estimated values.
2. A solution based on dynamic sliding windows for concept drift monitoring enables the effective updating of the missing value imputation model, resulting in increased estimated value accuracy.

Thus, as for development directions: 1) To address different missing data mechanisms (i.e., MCAR, MAR, MNAR) to perform imputations more consistent with real-

world problem characteristics, we will explore imputation solutions that consider relationships between attributes (Hypothesis 1). 2) To incrementally identify and adapt to different types of concept drift, we will implement dynamic window solutions (Hypothesis 2).

2. Methodology

The development of this research involves the following main steps: 1) Data collection and exploratory analysis; 2) Missing values impact assessment; 3) Missing values handler method conception and implementation; 4) Experimental tests with correlated approaches; 5) Results evaluation and discussion.

For the first step, publicly available datasets related to the healthcare domain will be used, mainly those obtained from wearable data sources. The work of Isgur *et al.* (2022) compiles public datasets with healthcare stream data (e.g., heart rates, steps, temperatures, sleep duration, ...) from patients across long timestamps on small granularities. We are currently working on the COVID-19 Wearables Datasets [Mishra et al. 2020], further presented in the next section, along with the results of the second step of our proposed methodology, the missing values impact assessment. This step is fundamental to quantitatively measuring the impact of missing values in data mining tasks for wearable data and motivating this research project.

The third step focuses on developing and implementing a method for handling missing values in data streams, which is the main contribution of this work. This solution aims to be more than just an algorithm for imputing missing values; it will provide a well-structured workflow with processing steps that meet all the proposed requirements. To achieve this, we will follow the development directions outlined in Section 1.2 while continuously reviewing the literature to stay updated with state-of-the-art solutions for future experiments and comparisons.

To experiment and validate our method, the fourth and fifth step, respectivaly, consists of comparative experiments that will be conducted using feasible correlated methods from the literature (Table 1) and evaluated with hypothesis tests comparing multiple performance metrics results, such as RMSE and MAE for imputatiuon quality and accuracy, precision, recall and F1 Score for classification task improvement assessment. The results will be discussed and made available in future publications, alongside the finalization of this PhD degree.

3. Preliminary results

To asses and measure the impact of missing values presence in healthcare wearable data mining task, we simulated missing values, considering each missing mechanics characteristics, in a healthcare dataset obtained from wearable devices of COVID-19 positive patients, presented in the work of Mishra *et al.* (2020). In the paper, an anomaly detection task was conducted using the Resting Heart Rate (RHR) metric, built upon each patient's heart rate and steps. Majorly, the anomalies indicated the presence of the disease in a pre-symptomatic period.

Within this context, we simulated missing values in the heart rate data from 27 patients with reproducible original results using the mdatagen Python library

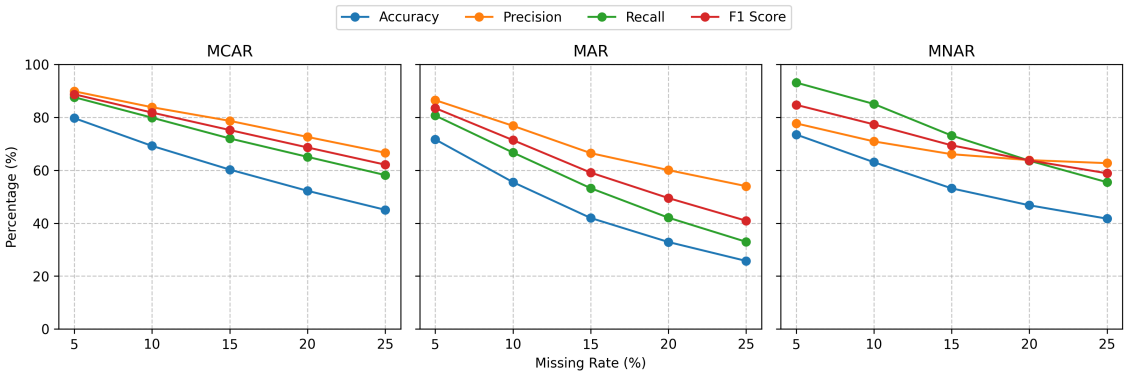

Figure 1. Performance of the accumulated results of the 27 patients for each missing rate and missing mechanism

[Mangussi et al. 2024] developed upon the work of Santos *et al.* (2019). We simulated missing values using the "heartrate" feature: for MCAR, values were removed randomly; for MAR, values were removed during early hours of the day (based on "datetime") to reflect typical rest periods; and for MNAR, the lowest heart rate values were removed, assuming they occur during rest and may impact RHR anomaly detection. For each mechanism, we simulated missing rates of 5%, 10%, 15%, 20%, and 25%, executing each 10 times for each patient. The results were compared with each patient's original ground-truth set of original anomalies, making it possible to measure the results quantitatively.

Figure 1 shows the performance of the accumulated results. Overall, the presence of missing values negatively affected all metrics, leading to poorer anomaly detection outcomes, even at the lowest missing rate of 5%. Among the metrics, accuracy exhibited the most substantial decline. When examining each mechanism, the Missing At Random (MAR) mechanism demonstrated the greatest performance drop across all missing rates, suggesting that missing values during the early hours of the day have a significant impact

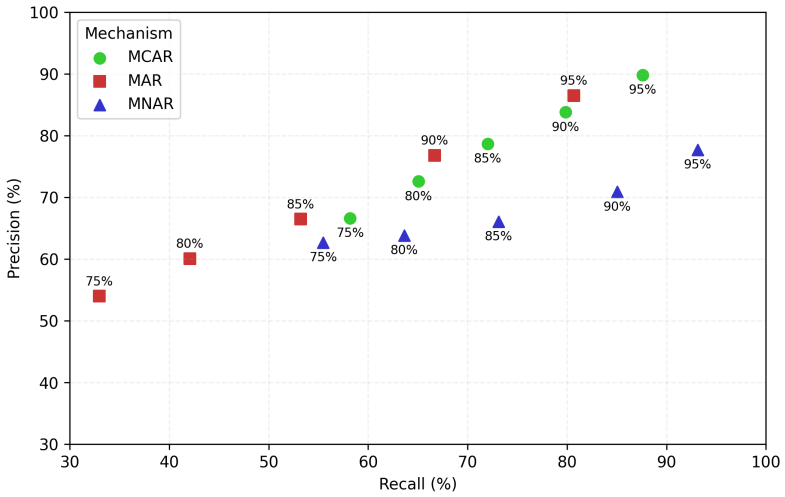

Figure 2. Recall X Precision of the accumulated results of the 27 patients for each missing rate and missing mechanism. The labels in each dot represent the percentage of the remaining original dataset (i.e., 100 - Missing Rate)

on the resting heart rate (RHR) metric. In contrast, the MNAR mechanism, with the lowest heart rate configuration, affected overall performance to a lesser extent than MAR, although its precision was the lowest among all mechanisms.

To better understand the performance impact of different mechanisms, Figure 2 illustrates the relationship between recall and precision results. For the MCAR and MAR mechanisms, both precision and recall remained similar across varying missing rates. However, the MAR mechanism showed a greater decline in both metrics compared to MCAR; for instance, at a 15% missing rate in MAR, the results were worse than those seen at a 25% missing rate in MCAR. In contrast, MNAR low missing rates resulted in better recall than other mechanisms, although it exhibited the poorest precision overall. This suggests that missing values in the MNAR mechanism do not significantly hinder the identification of the correct anomaly timestamp, as reflected in the high recall scores. However, the low precision indicates an overestimation of the anomaly period, leading to more false positives. Consequently, this may prolong the patient anomaly period and potentially affect the pre-symptomatic detection of diseases.

Furthermore, while the overall results indicate that the performance is unbiased, individual patients may be affected differently by the configuration of the missing mechanism. For example, Figure 3 shows the performance heatmap of a patient identified as "AIFDJZB." In this case, starting from a 10% missing rate under the MAR mechanism, all metrics returned 0% results, meaning that no anomalies were detected. Upon further analysis, we found that the MAR configuration, which caused values to be missing during the early hours of the day, led to the absence of critical heart rate data necessary for identifying patient anomalies, even with only a 10% missing rate. This case highlights the importance of each heart rate measurement and why it cannot be overlooked, motivating the use of more sophisticated imputation methods tailored to the MAR mechanism.

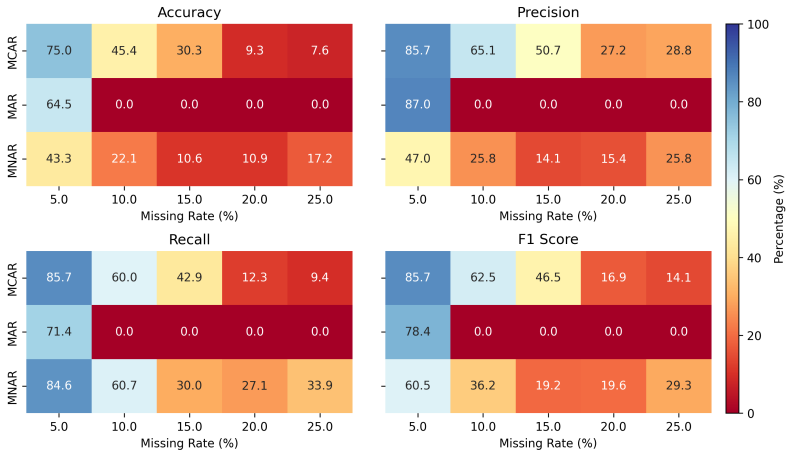

Figure 3. Heatmap with the performance of the accumulated results of patient labeled "AIFDJZB" for each missing rate and missing mechanism

References

Bahri, M., Bifet, A., Gama, J., Gomes, H. M., and Maniu, S. (2021). Data stream analysis: Foundations, major tasks and tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(3):e1405.

Dong, W., Gao, S., Yang, X., and Yu, H. (2021). An exploration of online missing value imputation in non-stationary data stream. *SN Computer Science*, 2:1–11.

Fountas, P. and Kolomvatsos, K. (2020). A continuous data imputation mechanism based on streams correlation. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE.

Getzen, E., Ungar, L., Mowery, D., Jiang, X., and Long, Q. (2023). Mining for equitable health: Assessing the impact of missing data in electronic health records. *Journal of biomedical informatics*, 139:104269.

Halder, B., Ahmed, M. M., Amagasa, T., Isa, N. A. M., Faisal, R. H., and Rahman, M. M. (2022). Missing information in imbalanced data stream: fuzzy adaptive imputation approach. *Applied Intelligence*, 52(5):5561–5583.

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.

Li, X., Li, H., Lu, H., Jensen, C. S., Pandey, V., and Markl, V. (2023). Missing value imputation for multi-attribute sensor data streams via message propagation. *Proceedings of the VLDB Endowment*, 17(3):345–358.

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, Porto Alegre, RS, Brasil. SBC.

Liu, W., Luo, L., and Zhou, L. (2023). Online missing value imputation for high-dimensional mixed-type data via generalized factor models. *Computational Statistics & Data Analysis*, 187:107822.

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.

Psychogyios, K., Ilias, L., Ntanos, C., and Askounis, D. (2023). Missing value imputation methods for electronic health records. *IEEE Access*, 11:21562–21574.

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.

Sun, Z., Zeng, G., and Ding, C. (2020). Imputation for missing items in a stream data based on gamma distribution. In *International Conference on Smart Computing and Communication*, pages 236–247. Springer.

Zhang, Y. and Thorburn, P. J. (2022). Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. *Future Generation Computer Systems*, 128:63–72.