# From Wearable Biometrics to Reproducible AI: Generalizing Evaluation Guidelines for Human-Centered Research

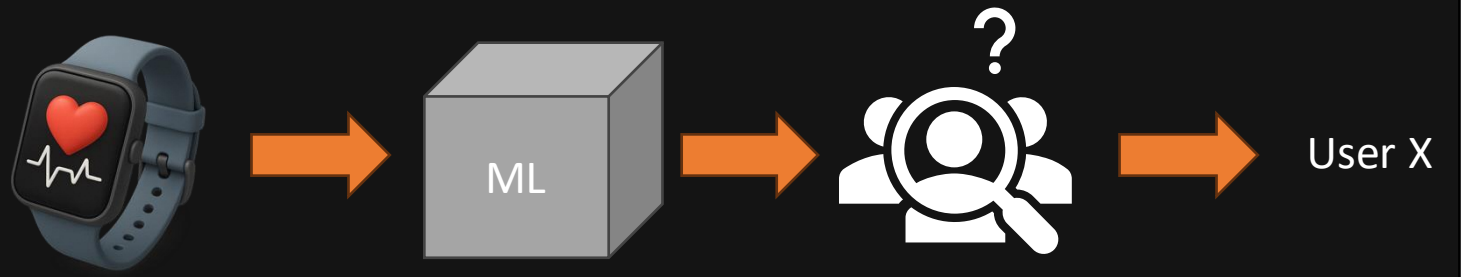
***Lidia Alecci***, Matías Laporte, Leonardo Alchieri, Nouran Abdalazim, Silvia Santini

Research on Research Event 2026 @ University of Zürich

29/01/2026

# Background

- Case study: focus on biometric recognition




- Why: different prior work shows very different performances

- Finding: differences in datasets and evaluation protocols affect final performance



*sensors*

MDPI

*Article*

## What the Heart Can(not) Tell: Potential and Pitfalls of Biometric Recognition Methods Based on Photoplethysmography

**Lidia Alecci, Matías Laporte, Leonardo Alchieri, Nouran Abdalazim and Silvia Santini ***

Faculty of Informatics, Università della Svizzera Italiana (USI), 6962 Lugano, Switzerland; lidia.alecci@usi.ch (L.A.); matiaslaporte@gmail.com (M.L.); leonardo.alchieri@usi.ch (L.A.); nouran.abdalazim@usi.ch (N.A.)
* Correspondence: silvia.santini@usi.ch

**Abstract**

Human physiological signals collected through wearable devices enable a range of applications, including biometric authentication. Prior studies have demonstrated the potential of using physiological signals to uniquely identify individuals, but their validity in real-world scenarios remains limited. Most existing work relies on controlled experimental settings, small datasets, short-term evaluations, and the absence of unseen-user testing—factors that tend to produce overly optimistic performance estimates. Although recent research highlights the need for broader benchmarking and reproducible protocols, systematic evaluations remain scarce. In this study, we assess the reliability of photoplethysmography (PPG)-based biometric methods. We replicate two published approaches and introduce a feature-based method as a baseline, evaluating all three under multiple conditions. Our results show that while these methods perform well in laboratory datasets, their effectiveness declines substantially in real-world environments, where signal variability, larger user populations, and temporal separation between training and testing challenge current systems. To address these issues, we propose guidelines for the robust evaluation of PPG-based biometrics, emphasizing real-world and longitudinal datasets, temporal splits, unseen-user assessments, and transparent reporting. Although developed for PPG, these recommendations generalize to other physiological biometrics and aim to improve the reliability and reproducibility of future research.

**Keywords:** photoplethysmography (PPG); biometric recognition; wearable sensors; real-world evaluation; user identification; evaluation guidelines

check for updates

Academic Editors: Wencheng Yang and Fei Zhu

Received: 14 November 2025
Revised: 9 December 2025
Accepted: 10 December 2025
Published: 14 December 2025

**Citation:** Alecci, L.; Laporte, M.; Alchieri, L.; Abdalazim, N.; Santini, S. What the Heart Can(not) Tell: Potential and Pitfalls of Biometric Recognition Methods Based on Photoplethysmography. *Sensors* **2025**, *25*, 7586. https://doi.org/10.3390/s25247586

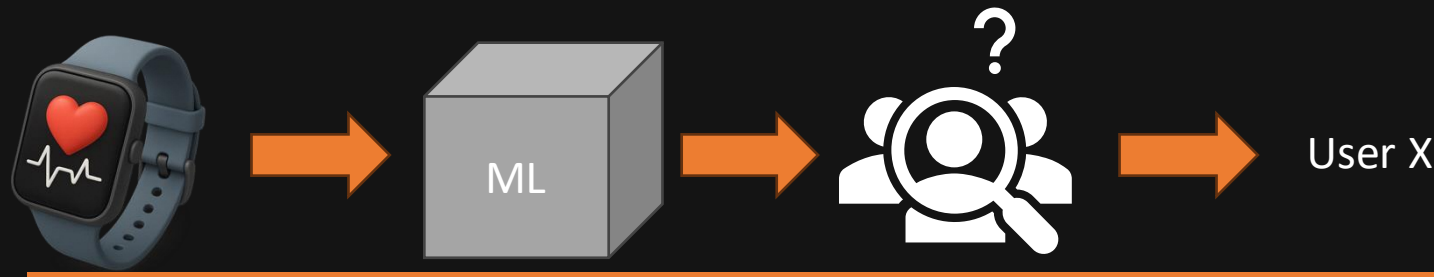


### 1. Introduction

Personal devices—such as smartphones, smartwatches, wearables, or smart rings—have become invaluable tools for collecting physiological data [1], including heart rate [2] and blood pressure [3,4] among others. These devices not only enhance health monitoring but also open new possibilities for authentication systems. Traditional methods, such as passwords and PINs, rely on knowledge-based credentials ("what you know"), making them vulnerable to issues like phishing and theft. In contrast, biometrics—the research field that studies how individuals can be uniquely recognized from their physical, chemical, or behavioral attributes [5]—rely on traits inherently possessed by individuals ("what you are"). Biometrics operates on the foundational assumption that unique patterns or characteristics can reliably distinguish one individual from another. These traits encompass facial features [6], fingerprints [7], and anatomical structures [8,9], as well

# Background

- Case study: focus on biometric recognition




Guidelines that extend beyond biometrics

- Finding: differences in datasets and evaluation protocols affect final performance



sensors

MDPI

Article

## What the Heart Can(not) Tell: Potential and Pitfalls of Biometric Recognition Methods Based on Photoplethysmography

Lidia Alecci, Matías Laporte, Leonardo Alchieri, Nouran Abdalazim and Silvia Santini *

Faculty of Informatics, Università della Svizzera Italiana (USI), 6962 Lugano, Switzerland; lidia.alecci@usi.ch (L.A.); matiaslaporte@gmail.com (M.L.); leonardo.alchieri@usi.ch (L.A.); nouran.abdalazim@usi.ch (N.A.)
* Correspondence: silvia.santini@usi.ch

**Abstract**

Human physiological signals collected through wearable devices enable a range of applications, including biometric authentication. Prior studies have demonstrated the potential of using physiological signals to uniquely identify individuals, but their validity in real-world scenarios remains limited. Most existing work relies on controlled experimental settings, small datasets, short-term evaluations, and the absence of unseen-user testing—factors that tend to produce overly optimistic performance estimates. Although recent research highlights the need for broader benchmarking and reproducible protocols, systematic

Accepted: 10 December 2025
Published: 14 December 2025

**Citation:** Alecci, L.; Laporte, M.; Alchieri, L.; Abdalazim, N.; Santini, S. What the Heart Can(not) Tell: Potential and Pitfalls of Biometric Recognition Methods Based on Photoplethysmography. *Sensors* **2025**, *25*, 7586. https://doi.org/10.3390/s25247586



sensors; real-world evaluation; user identification; evaluation guidelines

### 1. Introduction

Personal devices—such as smartphones, smartwatches, wearables, or smart rings—have become invaluable tools for collecting physiological data [1], including heart rate [2] and blood pressure [3,4] among others. These devices not only enhance health monitoring but also open new possibilities for authentication systems. Traditional methods, such as passwords and PINs, rely on knowledge-based credentials ("what you know"), making them vulnerable to issues like phishing and theft. In contrast, biometrics—the research field that studies how individuals can be uniquely recognized from their physical, chemical, or behavioral attributes [5]—rely on traits inherently possessed by individuals ("what you are"). Biometrics operates on the foundational assumption that unique patterns or characteristics can reliably distinguish one individual from another. These traits encompass facial features [6], fingerprints [7], and anatomical structures [8,9], as well

# Guidelines

**1. Reproducibility and transparency**

1.1 Open-source dataset

1.2 Open-source code

1.3 Detailed documentation


**2. Data quality**

2.1 Real-world data

2.2 Longitudinal data

2.3 Demographic diversity

2.4 Health condition diversity

2.5 Number of users

**3. Evaluation setup**

3.1 Temporal split

3.2 Testing with unseen users

3.3 Balanced evaluation

# Ensure reproducibility (code and data availability)

## Issue

Related work: **13**

with publicly available code: **1**

with publicly available data: **7**

## Solution

- Release code and datasets to support reproducibility
- Offer controlled data access (through data sharing agreements) when ethical constraints apply

# Ensure reproducibility (documentation)

## Issue

- “Random Forest was used”
  - No parameters or version specified


- “subtracting the moving average”
  - window size not specified

## Solution

- Always report library and version used
- Report all the necessary information

# Guidelines

**1. Reproducibility and transparency**

1.1 Open-source dataset

1.2 Open-source code

1.3 Detailed documentation


**2. Data quality**

2.1 Real-world data

2.2 Longitudinal data

2.3 Demographic diversity

2.4 Health condition diversity

2.5 Number of users

**3. Evaluation setup**

3.1 Temporal split

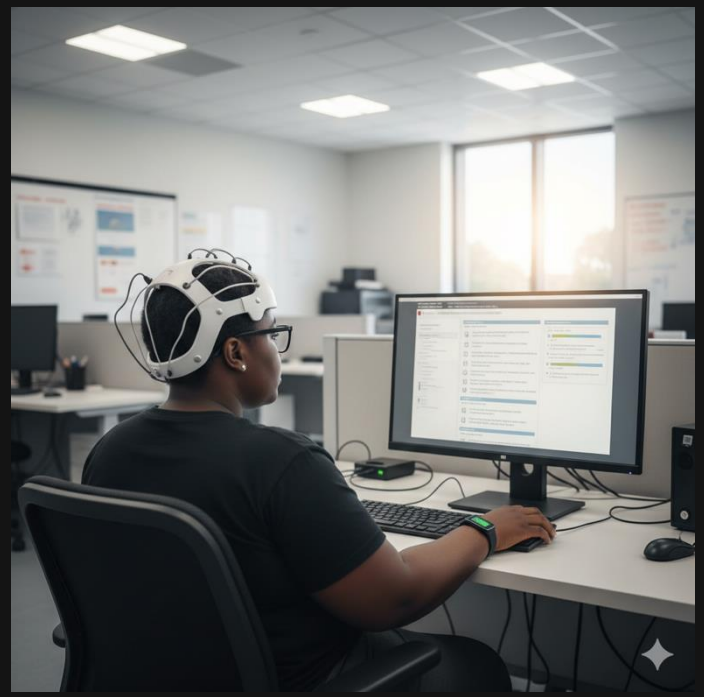
3.2 Testing with unseen users

3.3 Balanced evaluation

# Ensure data quality (laboratory vs real-world scenario)


## Issue

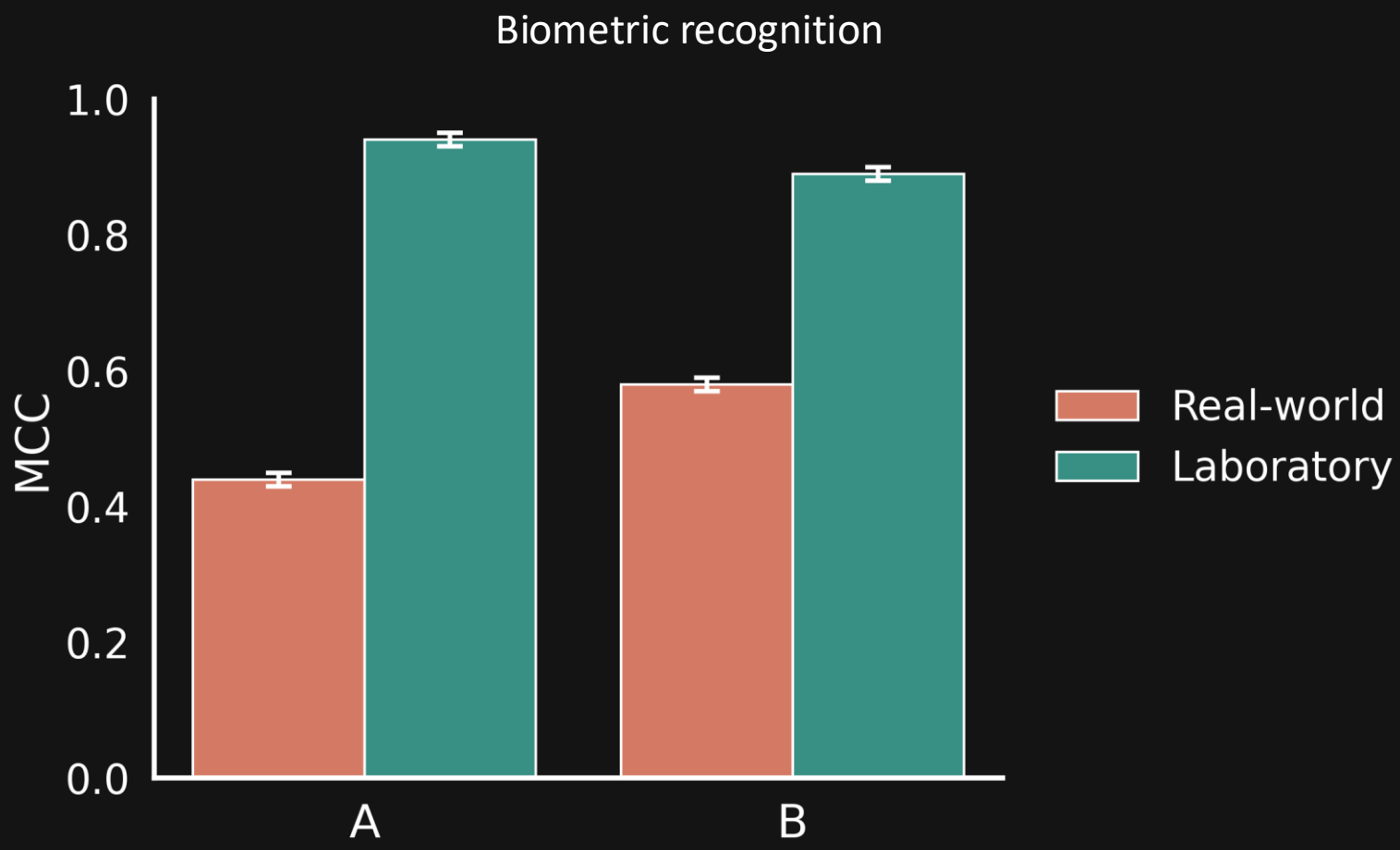
- Most of the solutions in literature are tested in laboratory settings



## Solution

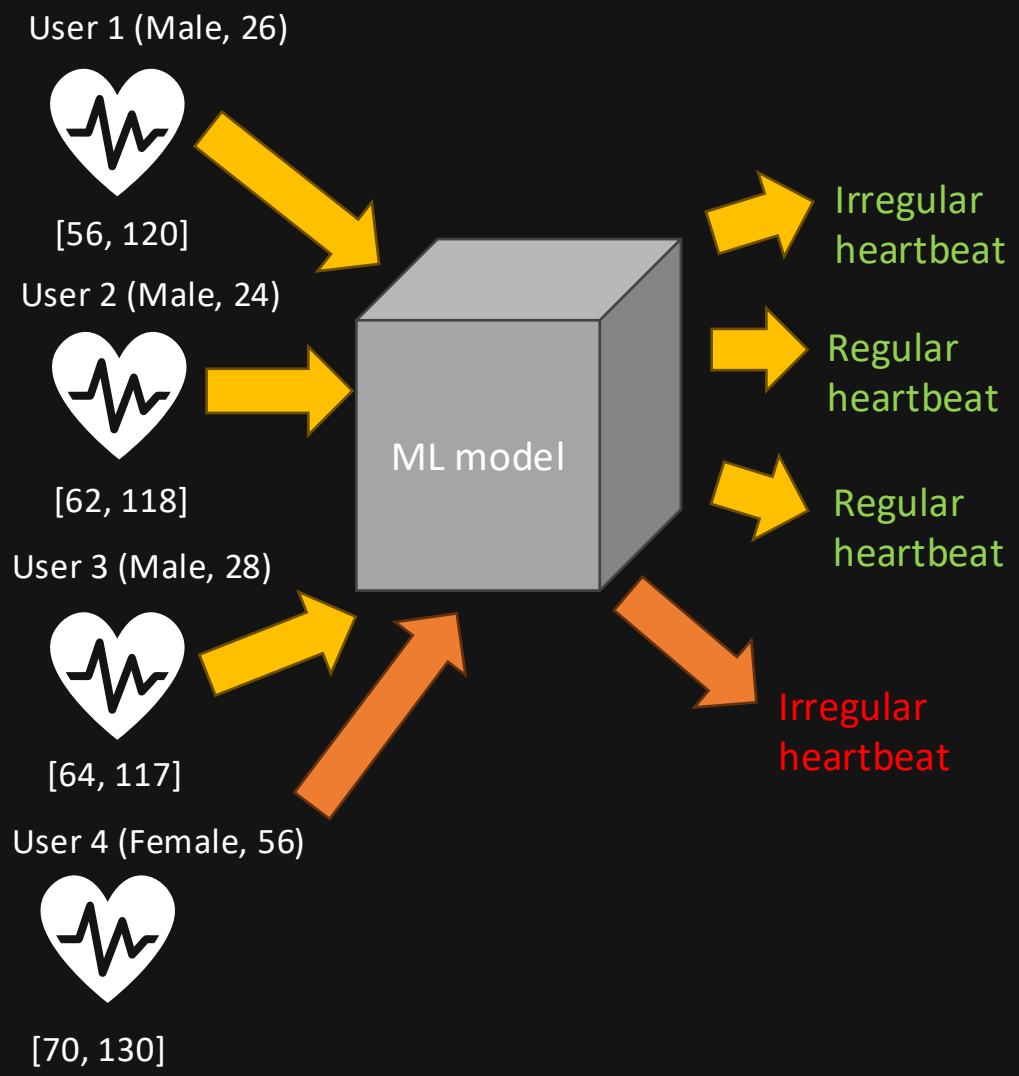
- Test solutions in real-world

# Laboratory vs real-world scenario – What we found



# Ensure data quality (diversity in the dataset)


## Issue




## Solution

- Demographic (e.g., age, gender) diversity
- Health-condition (e.g., chronic conditions) diversity

# Guidelines

**1. Reproducibility and transparency**

1.1 Open-source dataset

1.2 Open-source code

1.3 Detailed documentation


**2. Data quality**

2.1 Real-world data

2.2 Longitudinal data

2.3 Demographic diversity

2.4 Health condition diversity

2.5 Number of users

**3. Evaluation setup**

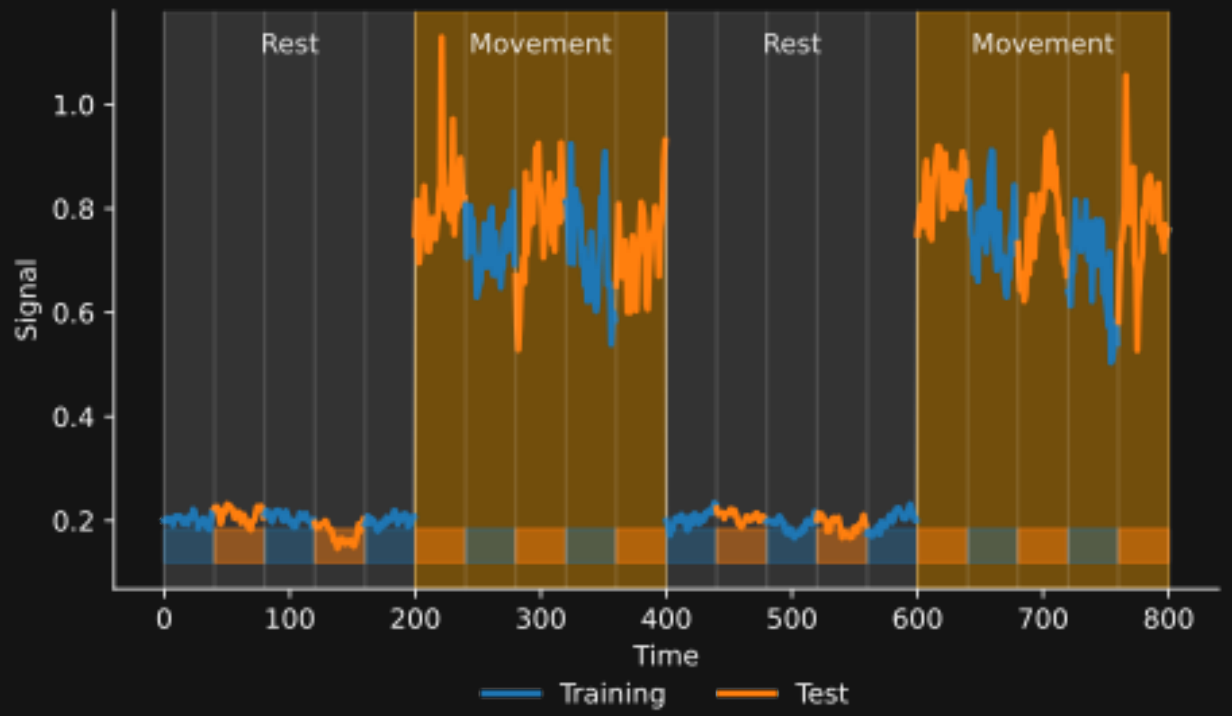
3.1 Temporal split

3.2 Testing with unseen users

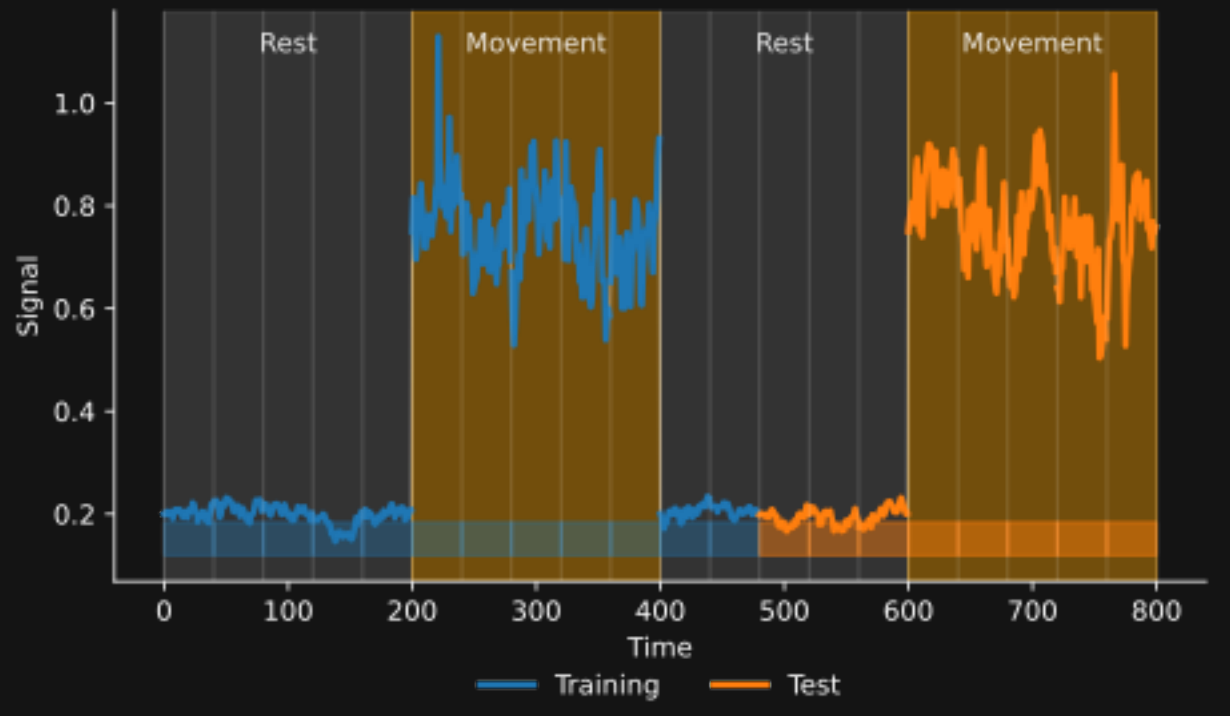
3.3 Balanced evaluation

# Ensure robust evaluation setup (training/test split)

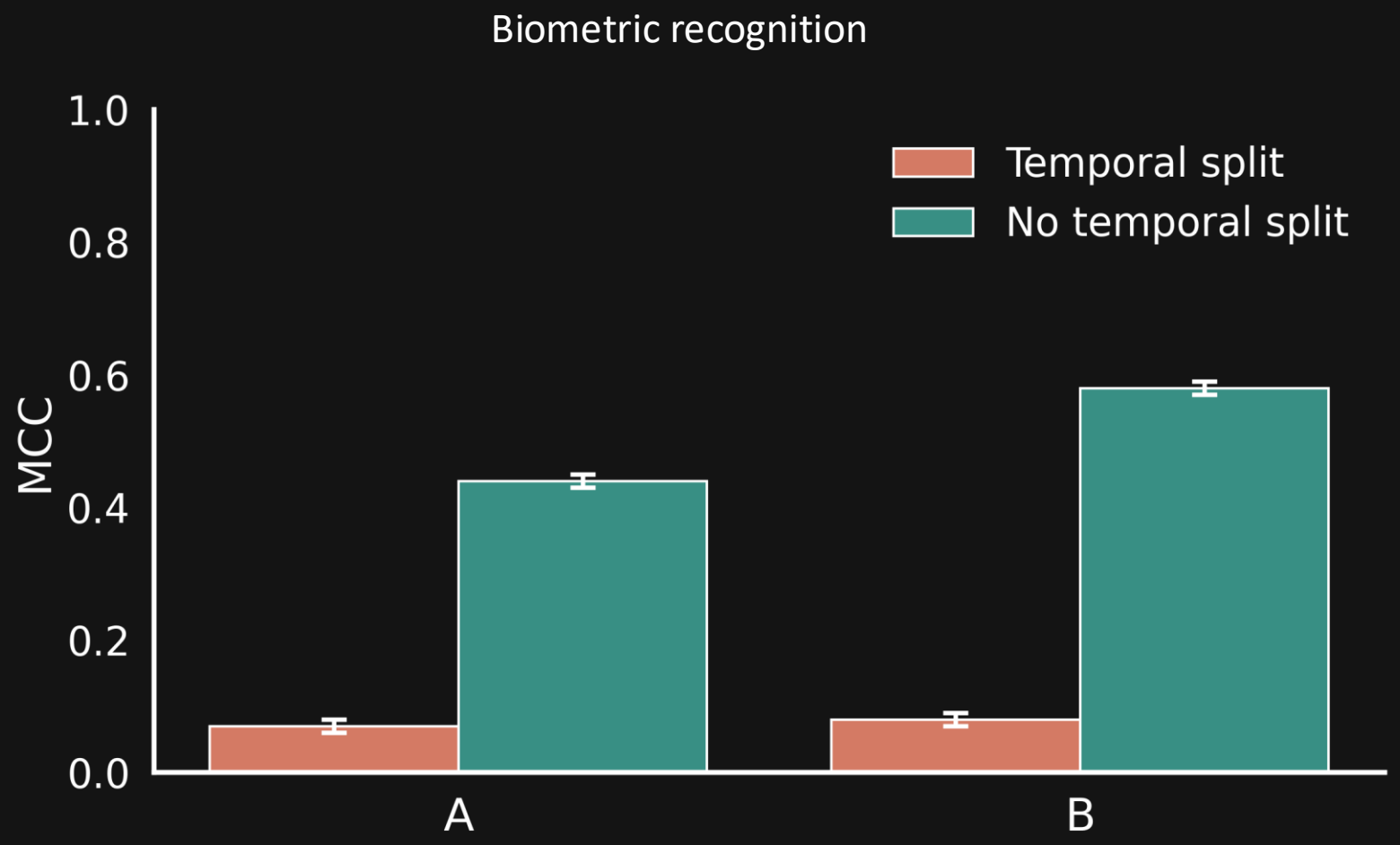
## Issue




## Solution



# Training/test split – What we found



# Towards Robust and Reproducible Evaluation

Paper

**1. Reproducibility and transparency**

1.1 Open-source dataset

1.2 Open-source code

1.3 Detailed documentation


**2. Data quality**

2.1 Real-world data

2.2 Longitudinal data

2.3 Demographic diversity

2.4 Health condition diversity

2.5 Number of users

**3. Evaluation setup**

3.1 Temporal split

3.2 Testing with unseen users

3.3 Balanced evaluation

Code repo

# Credits and references

- Icons taken from https://www.thiings.co
- Images (slide 7) generated using Nano Banana
- Alecci, L., Laporte, M., Alchieri, L., Abdalazim, N., & Santini, S. (2025). What the Heart Can(not) Tell: Potential and Pitfalls of Biometric Recognition Methods Based on Photoplethysmography. *Sensors*, *25*(24), 7586. https://doi.org/10.3390/s25247586
- Hammerla, N. Y., & Plötz, T. (2015). Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 1041–1051. https://doi.org/10.1145/2750858.2807551
- Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLOS Biology*, *13*(10), e1002264. https://doi.org/10.1371/journal.pbio.1002264
- Lara, O. D., & Labrador, M. A. (2013). A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys & Tutorials*, *15*(3), 1192–1209. https://doi.org/10.1109/SURV.2012.110112.00192