Assessing Multiple Imputation of Missing Values for Robust Analysis of Telehealth Kiosk Data

Hava Chaptoukaev, Maxime Beurey, Juliette Raffort, Maria A. Zuluaga

IA et santé : approches interdisciplinaires
June, 29^{th} 2022

Introduction

- **Telehealth** allows the distribution of health-related services.

- Promising avenue for **prevention**, and **remote diagnosis** and **monitoring** of diseases.

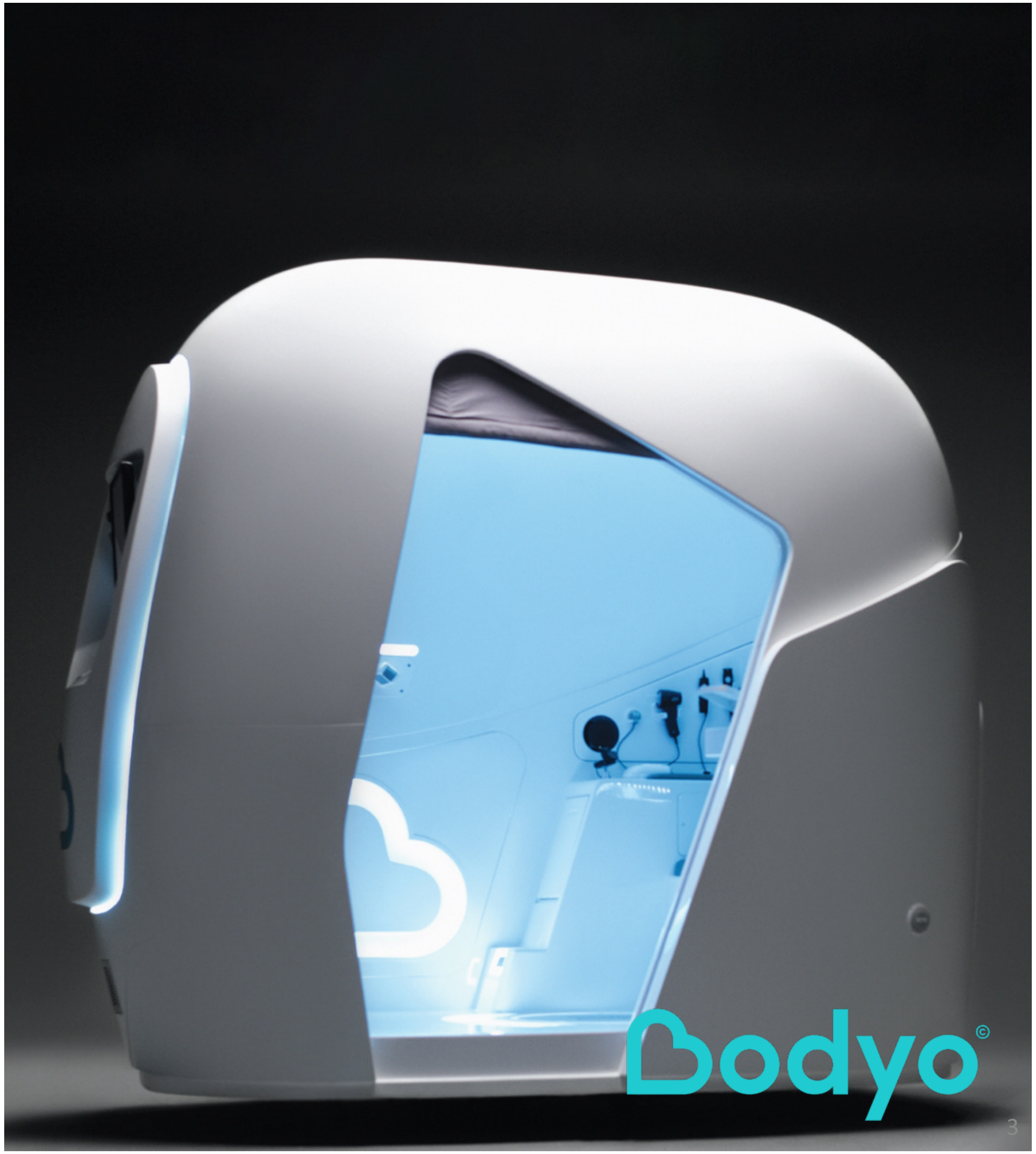
- Can be a solution when **access** to care is **restricted**.

EURECOM
Sophia Antipolis

Bodyo AiPod

- Stand-alone **telehealth** kiosk.
- Measures **27 health indicators** in 6 minutes.
- 4 **sensors** collect information :
  - a **scale**
  - a **body composition** sensor
  - an **oximeter**
  - a **blood pressure** sensor

EURECOM
Sophia Antipolis



Deploying the AiPod

- In non-clinical contexts **sensors may fail**, leading to incomplete data.
- If one sensor fails **all measures** collected by the sensor go missing.
- We cannot afford to **discard** incomplete observations.

Problem:
How to deal with missing data?

EURECOM
Sophia Antipolis

Bodyo©

Working with missing data

- We investigate two ways to deal with missing data:
  - **Imputation schemes** to fill the missing values.
  - A **set approach** that avoids **imputation.**

EURECOM
Sophia Antipolis

Bodyo©

Imputation of missing values

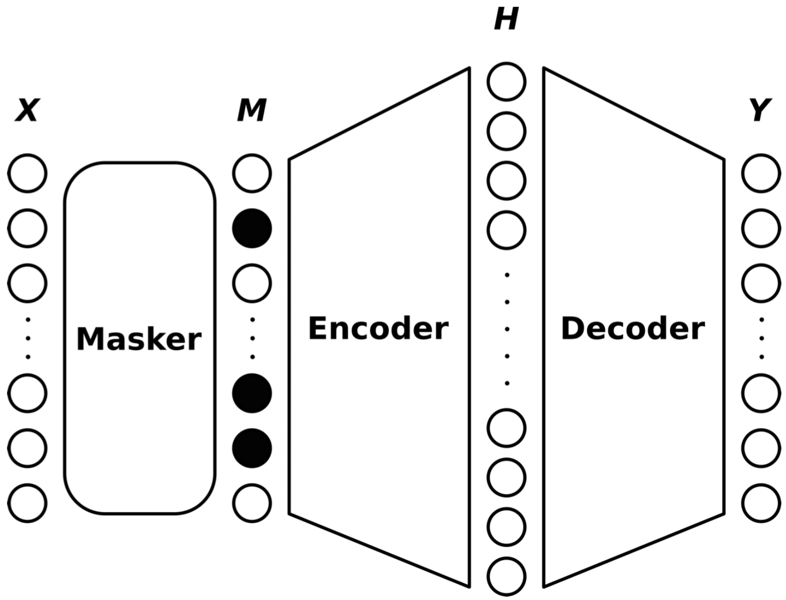
Imputation schemes

- We still want to **keep observations** with missing values.
- Imputation schemes allow to **fill** missing values.
- Imputation needs to **preserve the integrity** of the original data.

EURECOM
Sophia Antipolis

Multiple Imputation with Denoising Autoencoders (MIDA)

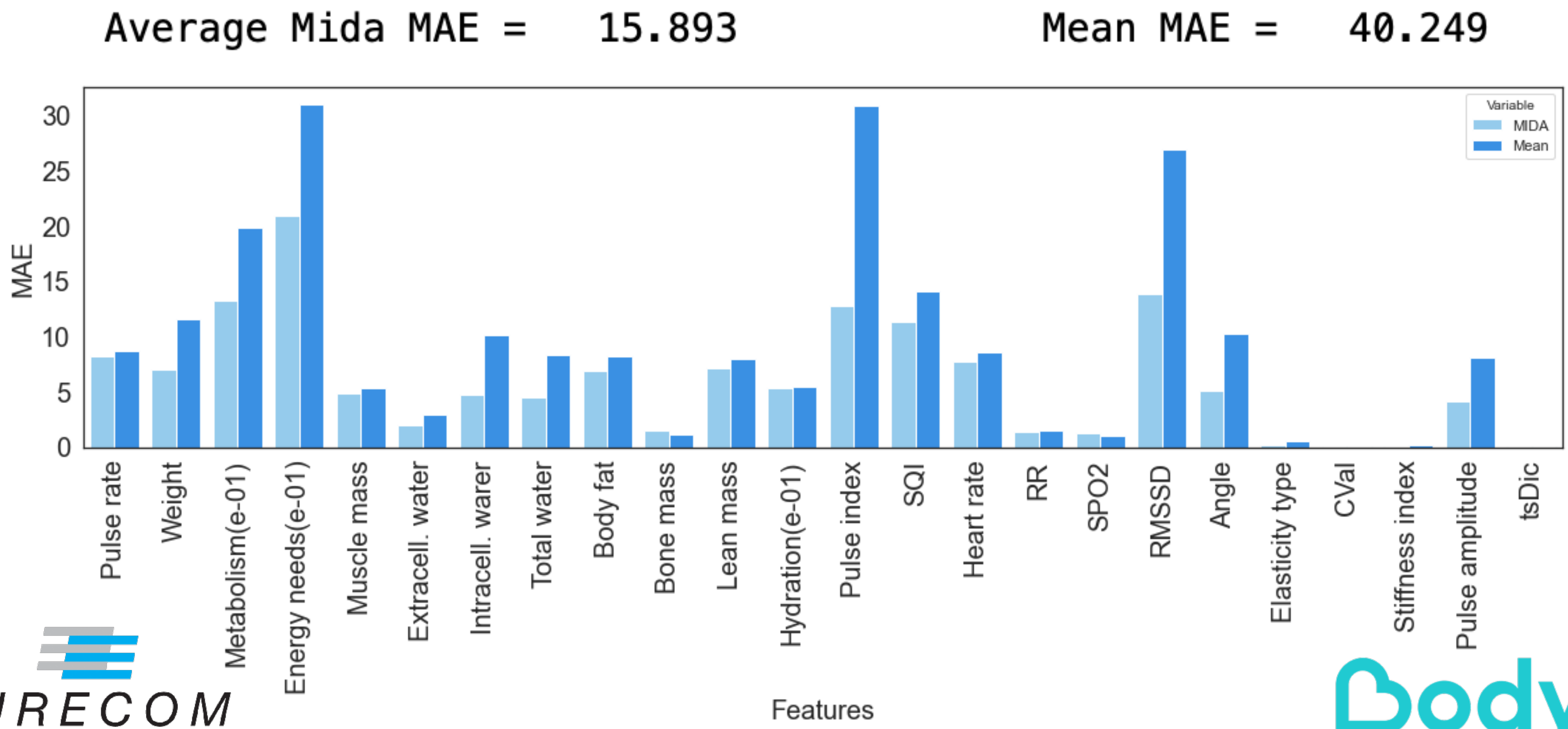
- The MIDA architecture[1] **imputes** missing values.
- Based on a **denoising** autoencoder.
- MIDA masker **not suitable** for our problem.



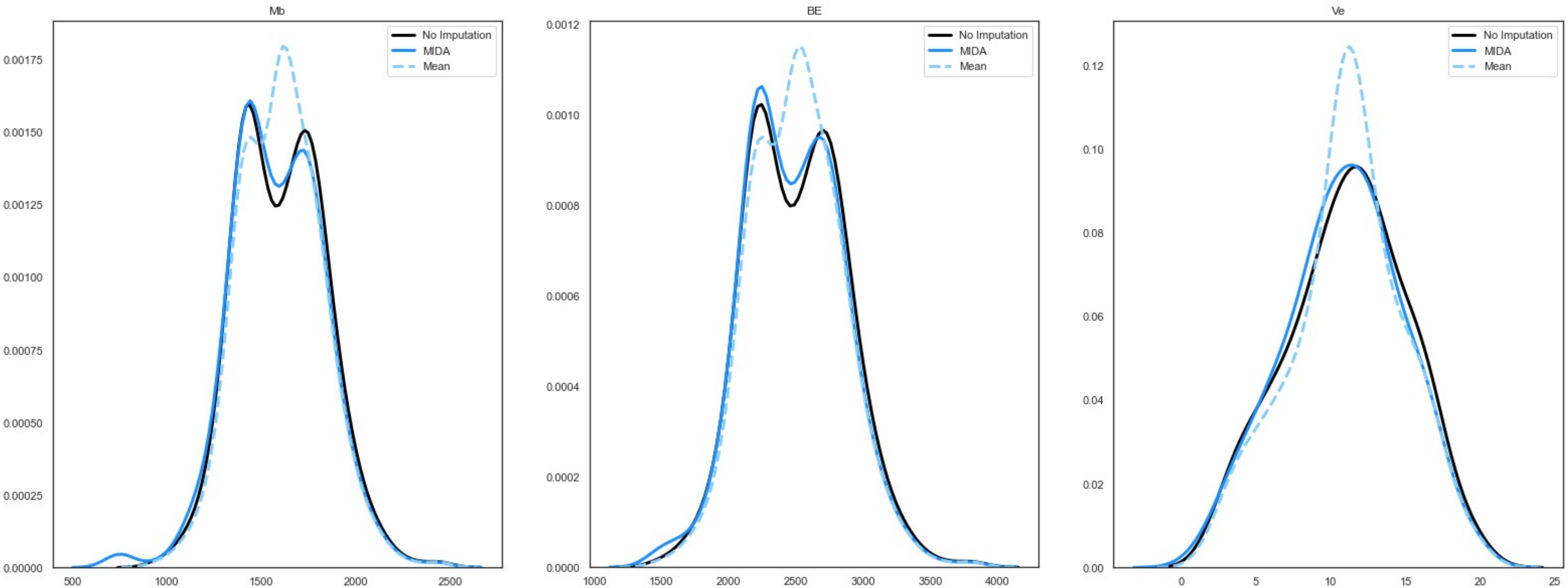
Our Approach:
Modify the masker to imitate the pattern of a failing sensor.

[1] Gondara, L., & Wang, K. (2018, June). Mida: Multiple imputation using denoising autoencoders. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 260-272). Springer, Cham.

Evaluation : feature reconstruction error



Evaluation : feature distribution



EURECOM
Sophia Antipolis

Bodyo©

Evaluation : blood pressure classification

- Dataset: **329 samples** with **24 features**.
- Data from at least one of the sensors is missing for **48 samples**.
- **Use case:** assess if imputation improves the binary classification of BP categories according to 2 categories.

	Accuracy	F1-score	Precision	Sensitivity
None	0.67	0.65	0.62	0.69
Mean	0.64	0.67	0.59	**0.77**
Our method	**0.71**	**0.71**	**0.65**	**0.77**

EURECOM
Sophia Antipolis

Bodyo©

Limitations of imputation

- There may be additional information, not collected by the sensors.
- In our dataset, only **105 samples** (out of 329) have no missing values.
- Imputation of poorly represented information can introduce **significant biases** in the learning process.

Set approach to learn with missing values

Set models

- Most classical machine learning models require **fixed-dimensional** inputs.

- **Sets** allow to overcome this limitation.

- Good **alternative** to learn with missing values.

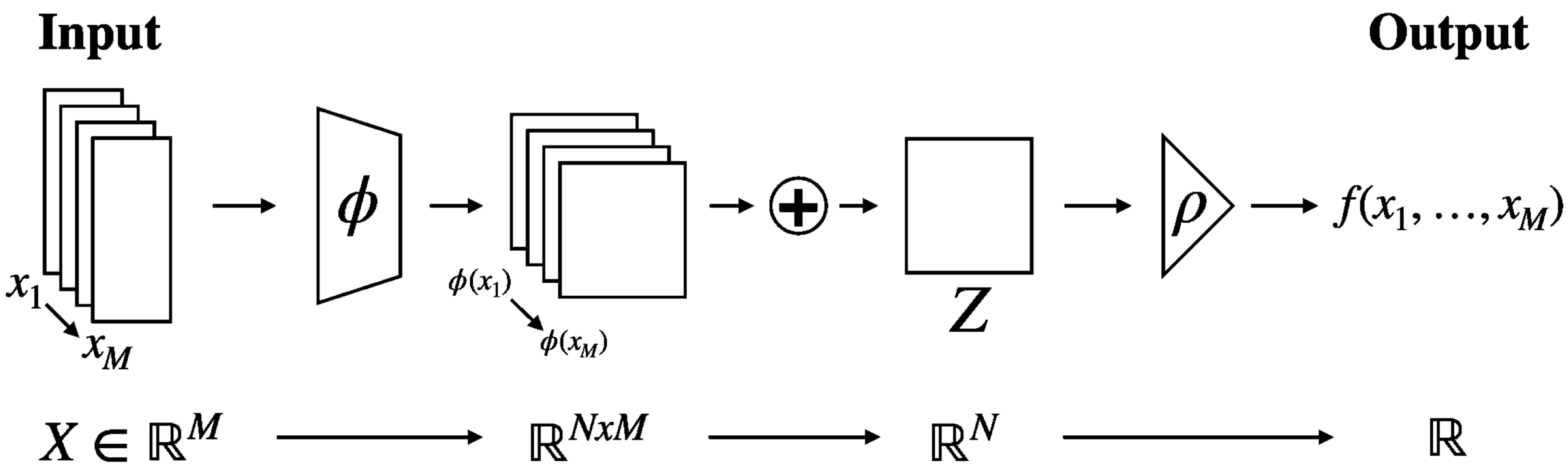
Set models

- **Idea**: use **permutation invariant** neural networks.
- **Permutation invariant function**: indifferent to the ordering of its input.

Theorem 1. *A function f operating on a set X having elements in a countable universe, is a valid set function iff there are functions $\varphi : R \to Z$ and $\rho : Z \to R$ such that*

$$f(X) = \rho\Big(\sum_{x \in X} \varphi(x)\Big) \tag{1}$$

Deep Sets architecture



EURECOM Sophia Antipolis

[2] Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., & Smola, A. J. (2017). Deep sets. *Advances in neural information processing systems, 30.*

Bodyo©

Our approach

- Each input vector X_i is **encoded as a set** of permutation invariant observations x_j.
- Each x_j is represented as a **tuple** $(v_j,\ m_j)$ such that :

$$X_i := \{(v_1,\ m_1),\ldots,(v_p,\ m_p)\}$$

- The whole dataset can then be described as:

$$D := \{(X_1,\ y_1),\ldots,(X_n,\ y_n)\}$$

Our Approach:
Allows us to deal with missing values.

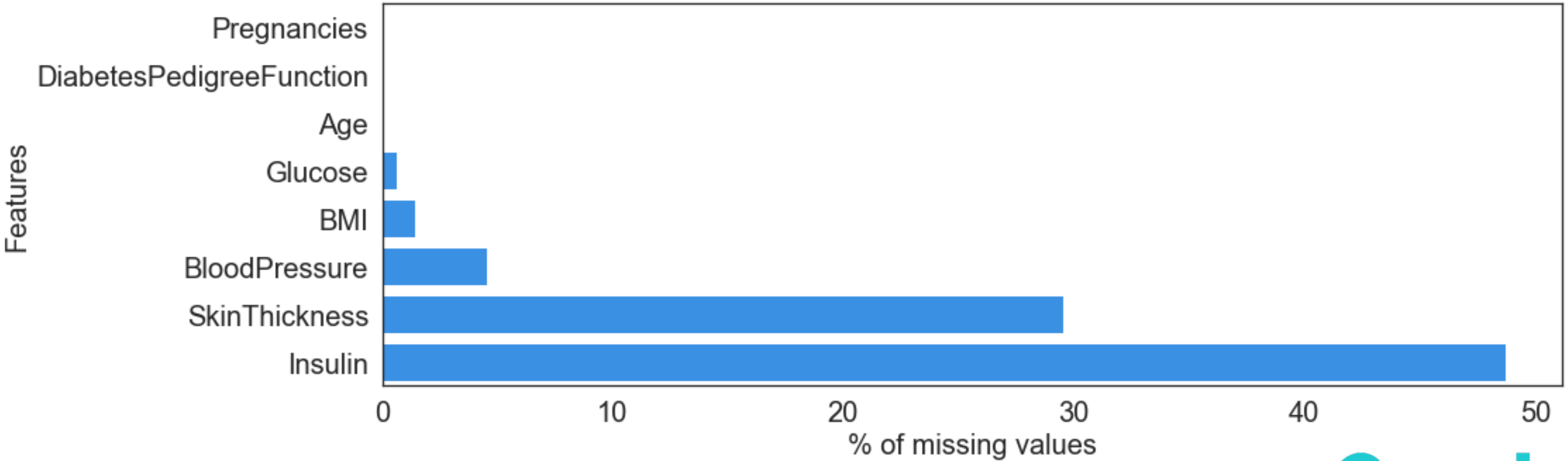
Paper in writting.

EURECOM
Sophia Antipolis

Bodyo©

Evaluation: diabetes classification

- The **Pima Indians Diabetes**[3] database is composed of **768 samples** and **8 features**.
- Up to **374 samples** have missing values across 5 features.



EURECOM Sophia Antipolis

[3] https://www.kaggle.com/competitions/diabetes-classification/data

Bodyo©

Evaluation : benchmark

- **Benchmark**: Logistic regression, Random Forests and Gradient Boosting.

- Missing values need to be **imputed** first for the benchmark.

	Accuracy	F1-score	Precision	Sensitivity
Mean imp. + LR	0.753	0.61	**0.70**	0.52
Mean imp. + RF	0.772	0.65	0.71	0.59
Mean imp. + GB	0.727	0.59	0.62	0.56
Our method	**0.792**	**0.71**	0.68	**0.74**

Concluding remarks

- The problem of missing values is a **particularly sensitive** issue in the medical field.
- We proposed two **simple** yet **robust** models that yield good performances.
- Imputation methods should be used **sparingly** to avoid biases in the learning.

EURECOM
Sophia Antipolis

Bodyo©

Ongoing work

- Develop a way to compute a **weighted aggregation**.
- Test the method on the **AiPod data.**
- Investigate the **combination** of the two approaches.

Thank you!

Assessing Multiple Imputation of Missing Values for Robust Analysis of Telehealth Kiosk Data

Hava Chaptoukaev, Maxime Beurey, Juliette Raffort, Maria A. Zuluaga

IA et santé : approches interdisciplinaires