



AI-based complication detection in neonatal intensive care data

Camelia Lauterbach Oprea, RWTH Aachen

Patient group: neonates (NICU)



<https://pageflow.gesundheitsforschung-bmbf.de/ki-fr-frhchen>

Treatment -> data, data, data

Over 500 pediatric patients
>250 neonatal patients
> 120 preterms

Mechanical ventilation:
airway flow, pressure,
FiO2, minute ventilation

Apnea bradycardia syndrome

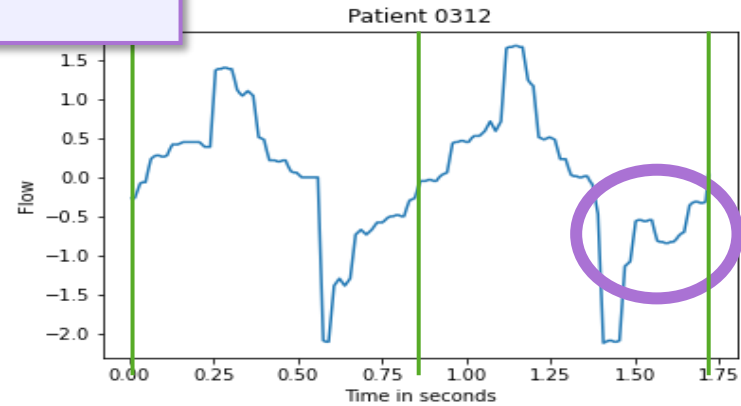
Patient ventilator asynchrony

Sensors: temperature, SpO2,
tcpCO2, ECG, chest impedance ...

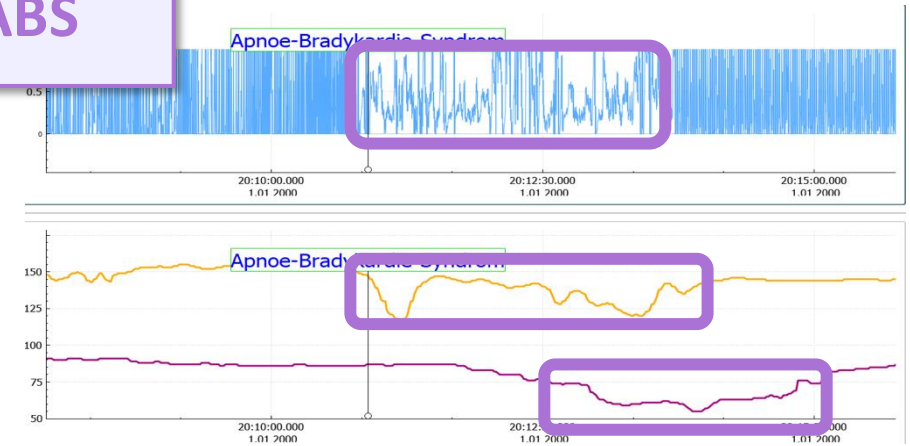
Olivier, L., Lauterbach, C. et al. "Feasibility of a German High-Frequency Neonatal and Pediatric Intensive Care Dataset (Alx-Neo-Guard Dataset): Secondary Data Cohort Study." *JMIR pediatrics and parenting* 9.1 (2026): e94280.

Complications

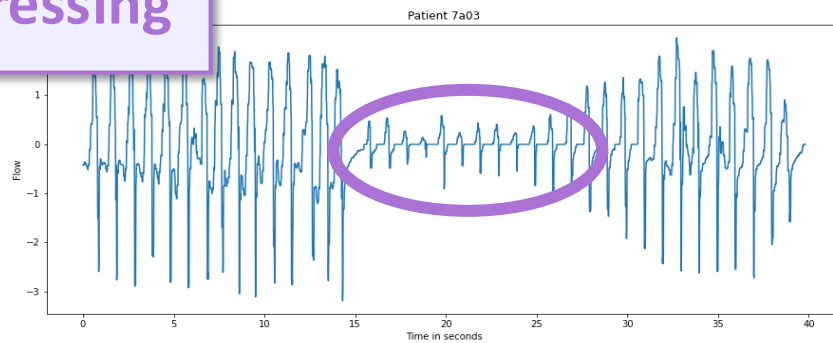
IEE



ABS

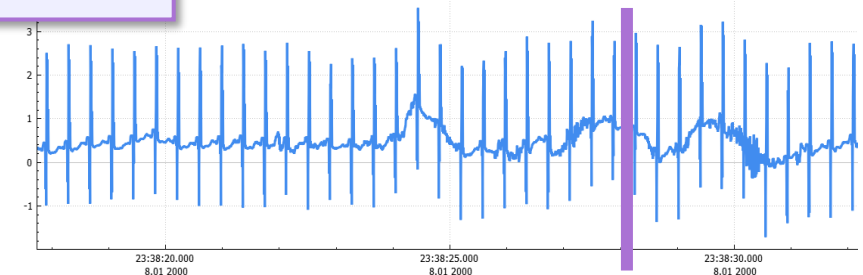


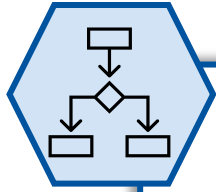
Pressing



Glucose

Glycemia: 100mg/dl





Algorithms



Use case
NICU



*Which **algorithms** are suited for an **AI-based complication detection** for **neonatal intensive care**?*

Constraints: *small labeled datasets, noisy data, high-frequency time series (waveforms)*

Algorithm development

Medical input

Idea, hypothesis
Definition of complication
Label data

Algorithm development

Data preprocessing
Feature extraction
Model choice
Training & validation

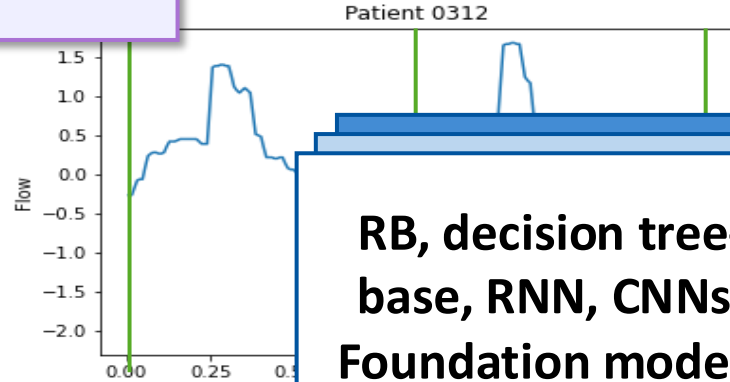
Evaluation

Test on unseen patient
Compare to state-of-the-art
External validation [1]

[1] Marconi, Luca, and Federico Cabitza. "Show and tell: A critical review on robustness and uncertainty for a more responsible medical AI." *International journal of medical informatics* 202 (2025): 105970.

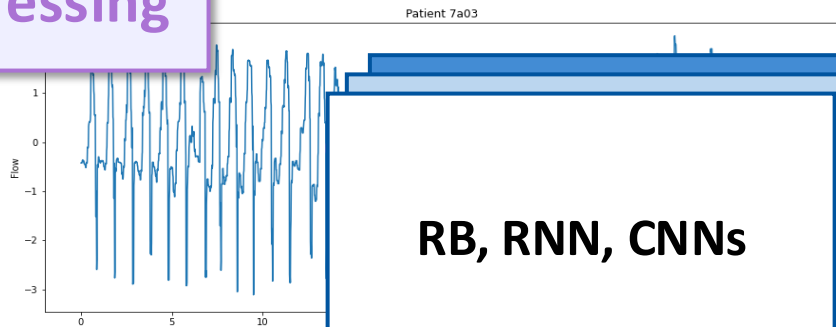
Complications – algorithms

IEE



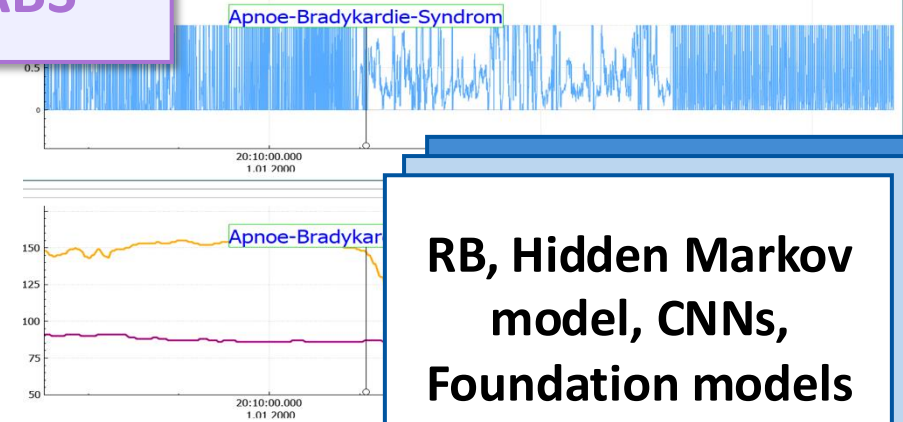
**RB, decision tree-
base, RNN, CNNs,
Foundation models**

Pressing



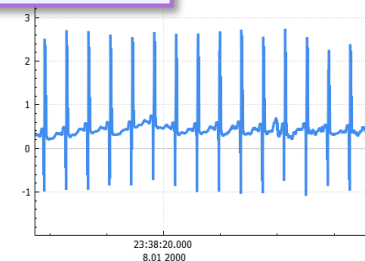
RB, RNN, CNNs

ABS



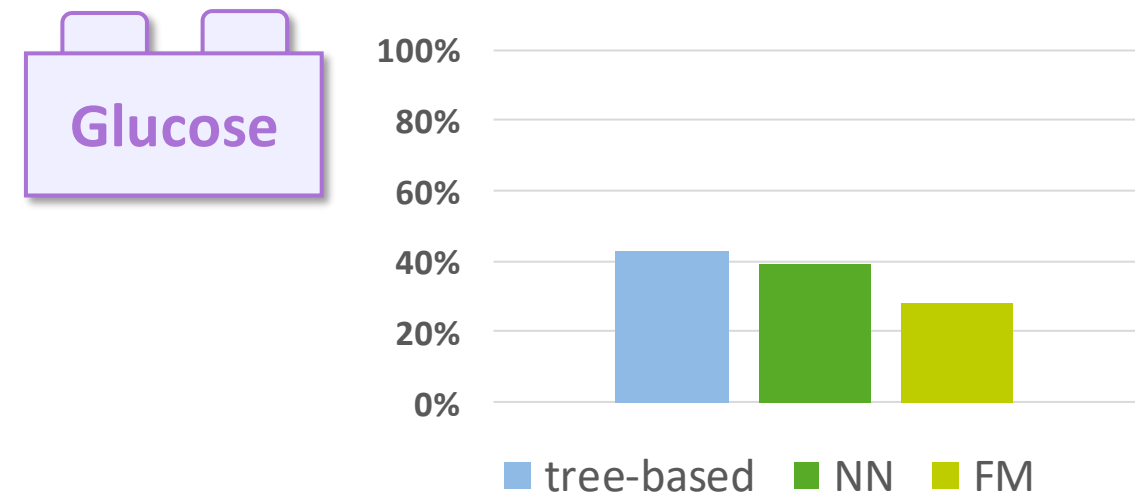
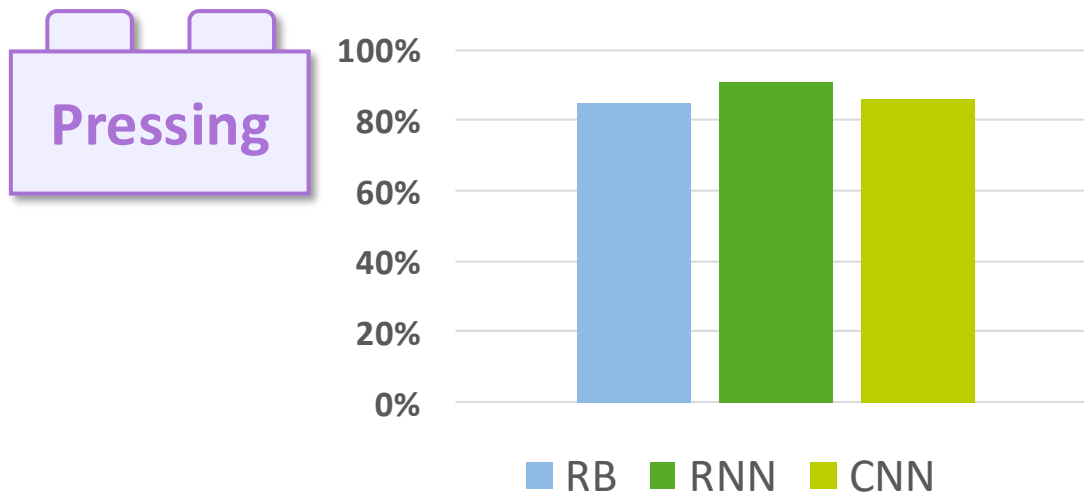
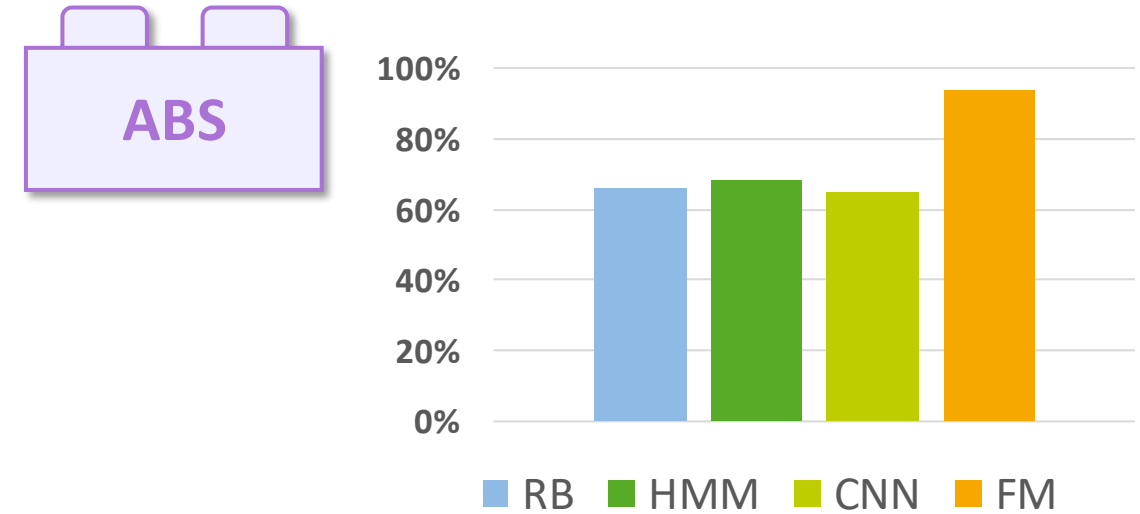
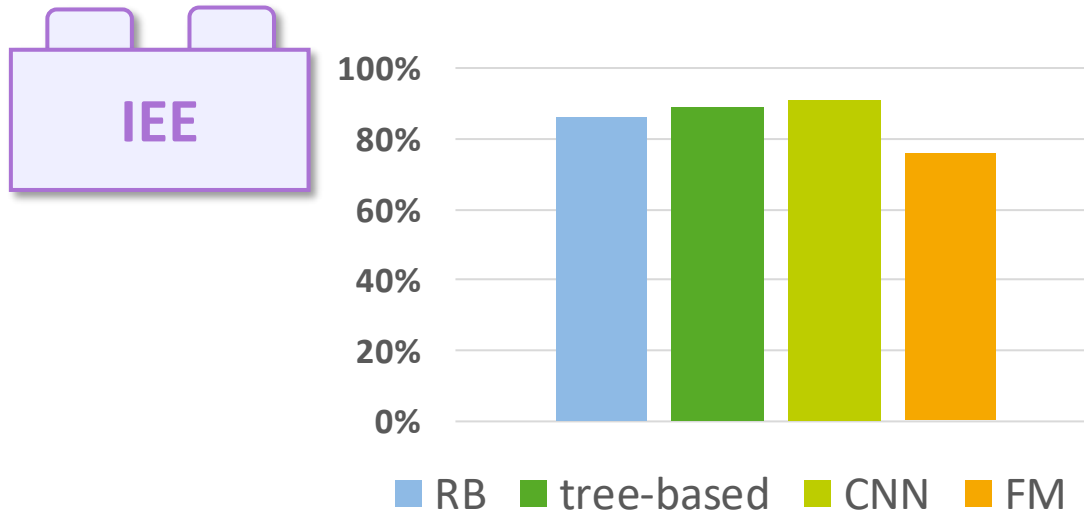
**RB, Hidden Markov
model, CNNs,
Foundation models**

Glucose



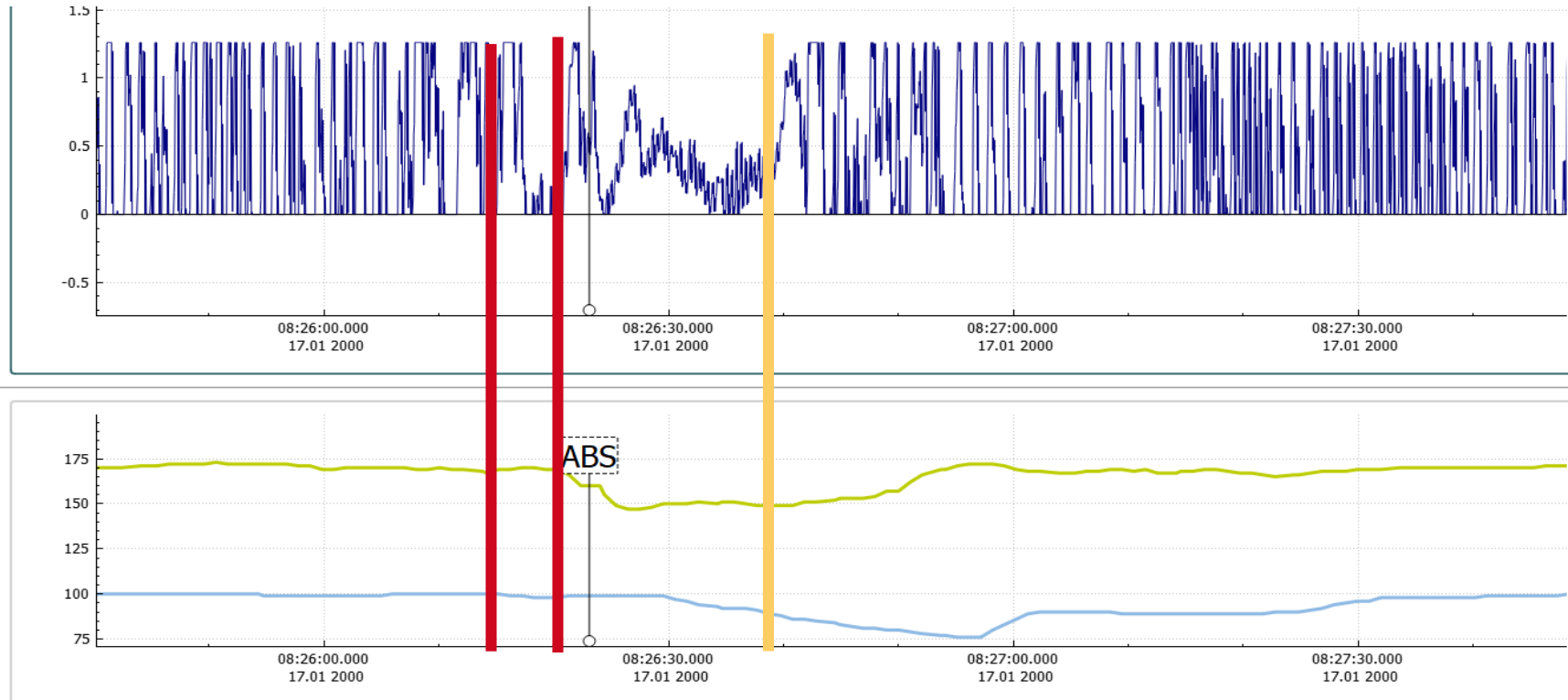
**Decision tree-based,
NNs, Foundation
models**

Complications – results (F1 score)



Labeling inconsistencies

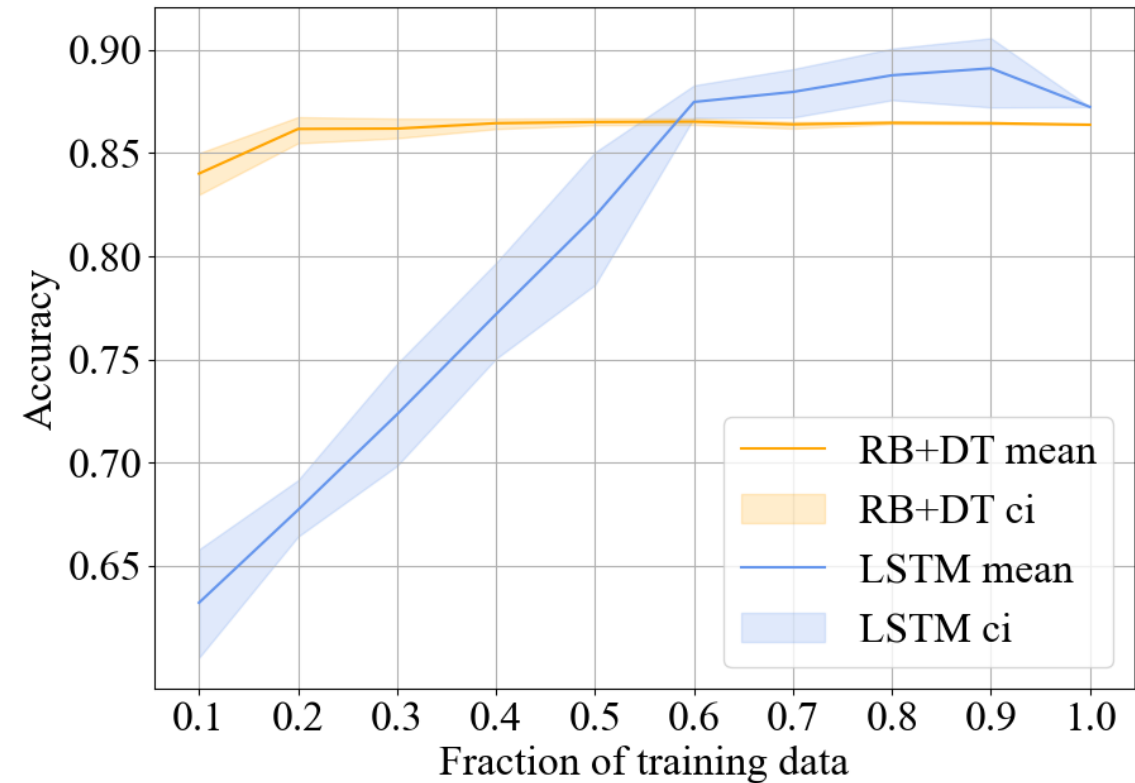
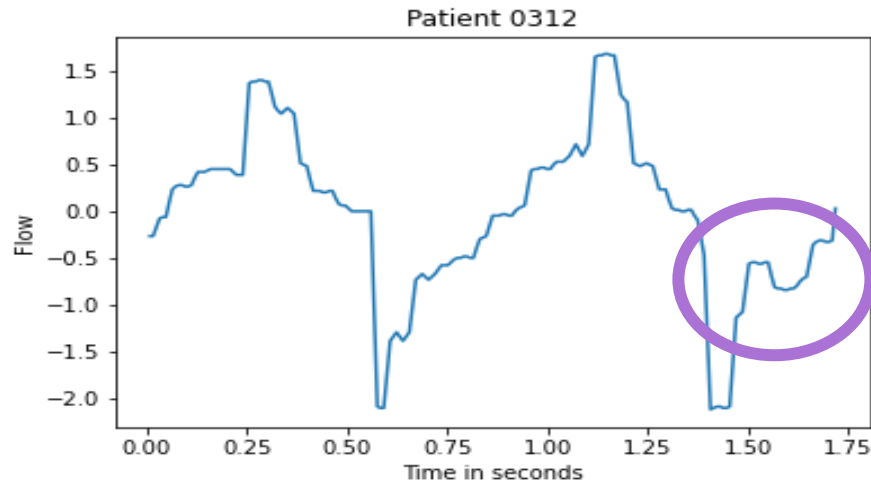
- ▶ Understanding varies across clinicians
- ▶ Labeling strategy differs (where to place annotation and how to describe it)
 - Label timing not exact
- ▶ Inconsistent



[2] Wei, Yishu, et al.
Journal of the American Medical Association 327.1 (2022): 1000-1007.

Effect of data amount

- ▶ Ineffective efforts during expiration
- ▶ Effect of training sample size on rule-based vs LSTM



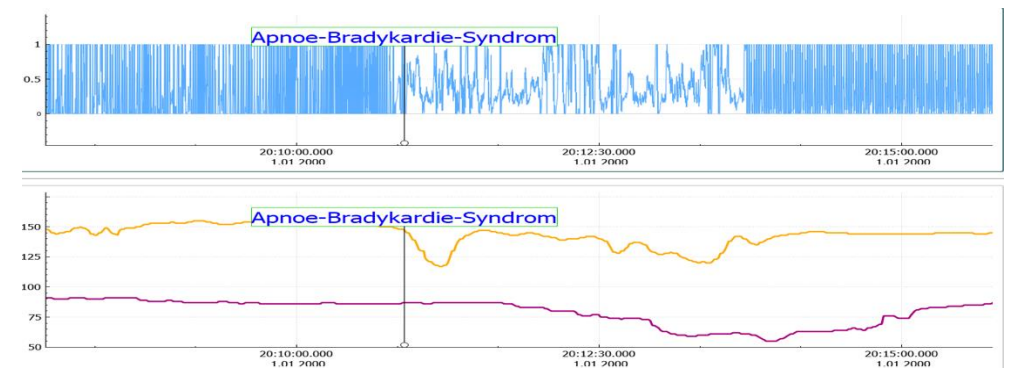
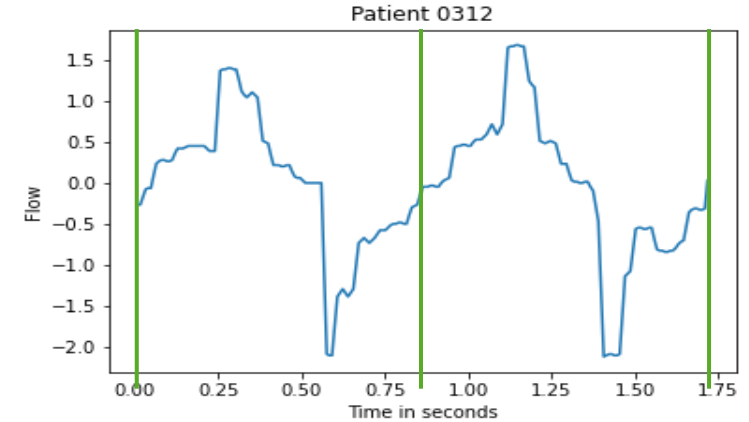
Foundation models – a solution for everything?

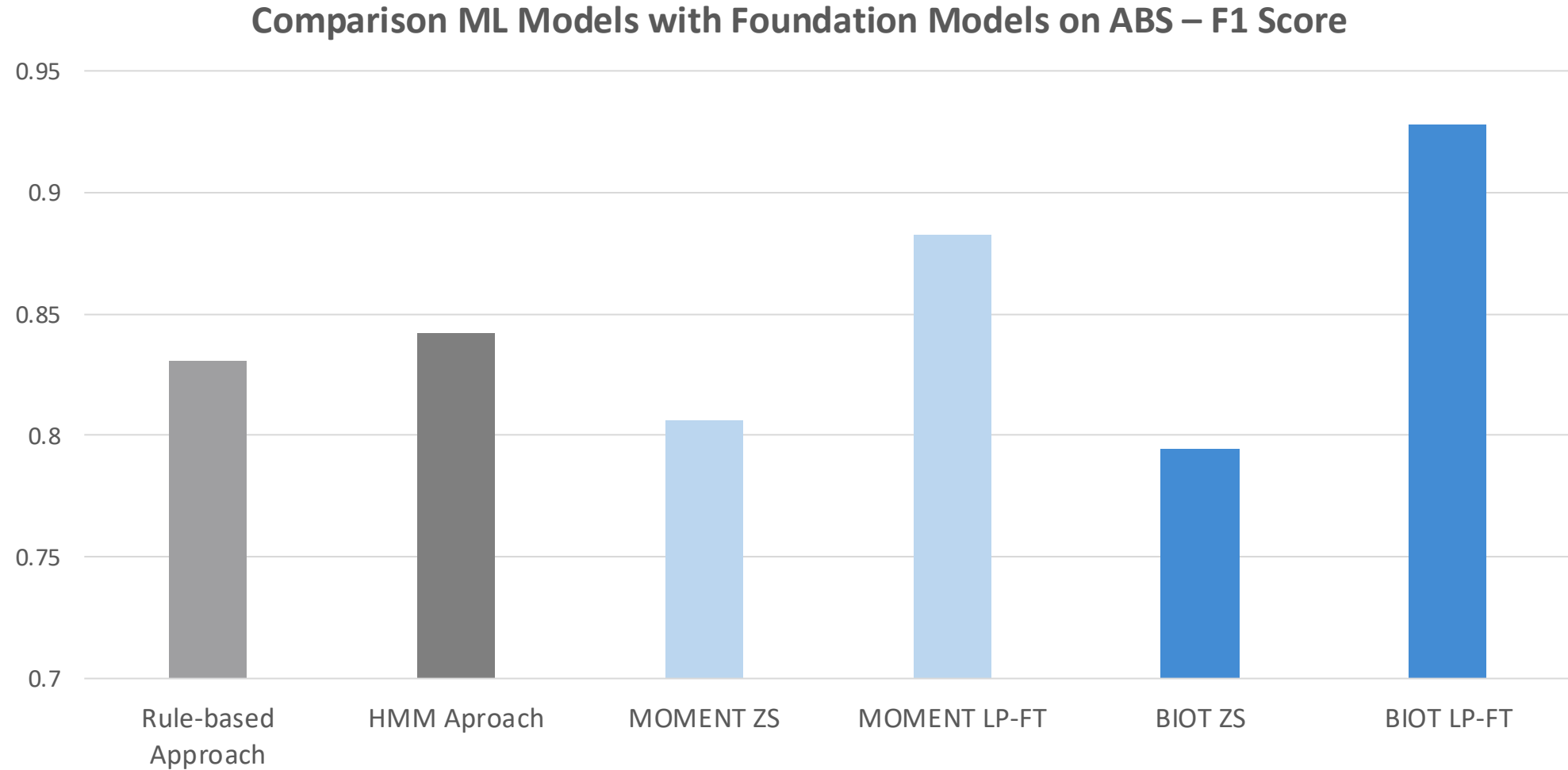
- ▶ Great for few labels

IEE (univariate) and **ABS** (multivariate)

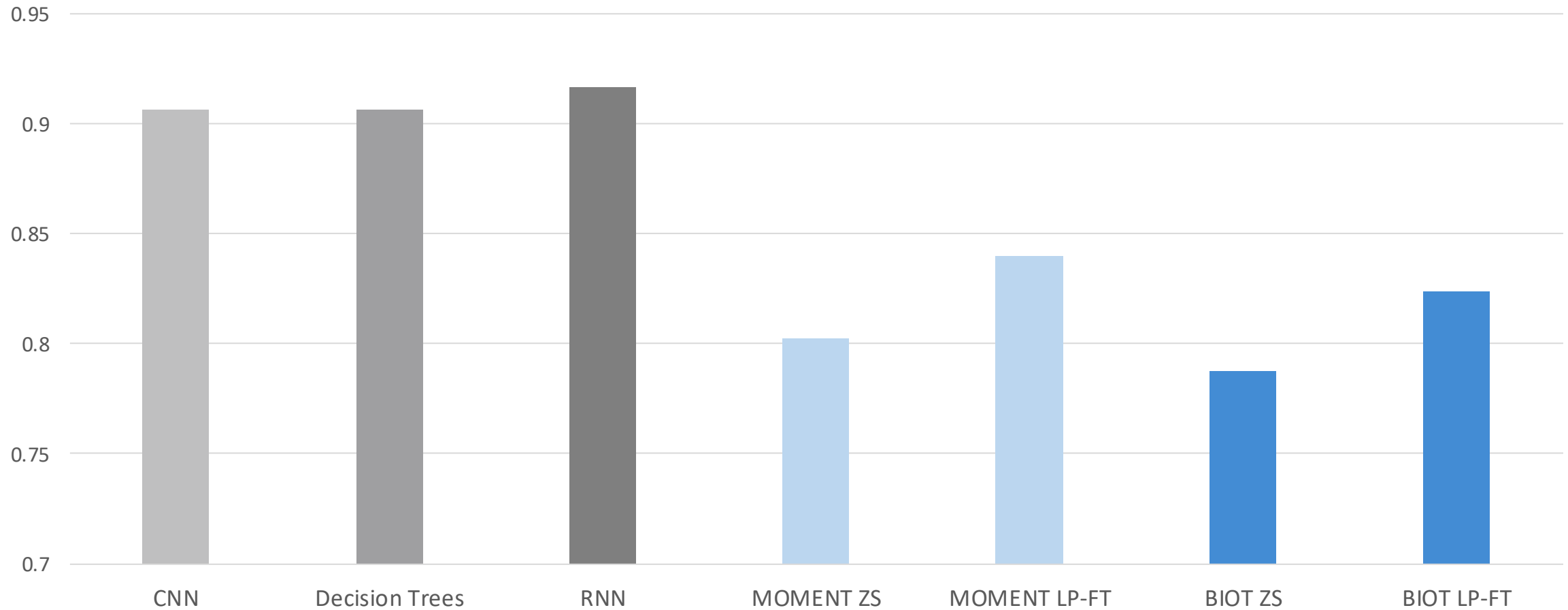
- ▶ Adapted two foundation models:

- BIOT (specialized on medical data)
- MOMENT
- MIRA, MOIRAI, AURORA.. (only for forecasting)
- Adaptation techniques: zero-shot, fine-tuning, linear probing, LP-FT
- Post-training on our own data





Comparison ML Models with Foundation Models on IEE – F1 Score



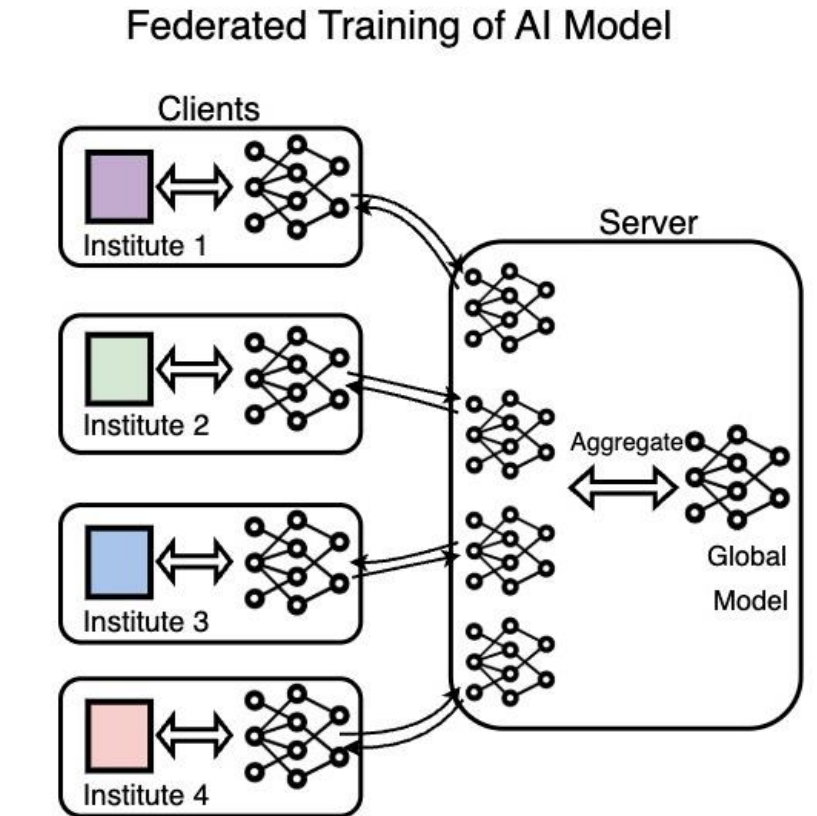
Glucose detection from ECG

- ▶ Self-supervised model pre-trained on 9.1 million ECGs
- ▶ Inspired by HuBERT transformer encoder for speech representation
- ▶ Evaluated on 21 datasets including 1.6k diagnostic targets
- ▶ Fine-tuned on glucose classification by adding output layer
- ▶ Best macro F1: **0.28** on 3-class classification (random guessing: **0.33**)
- ▶ Best macro F1 for self-trained model: **0.43**

[10] Coppola et al. (2024). HuBERT-ECG: a self-supervised foundation model for broad and scalable cardiac applications.

Federated learning – a solution for small datasets?

- ▶ Yes and no
- ▶ Few clinics save the data at high frequency
 - Very few save waveforms
- ▶ MIMIC Waveforms [3]
 - Only 6 neonates with waveform and clinical data
- ▶ More promising: transfer learning
 - Pretrain on adult data, if available



[3] Moody, Benjamin, et al. "MIMIC-III Waveform Database Matched Subset" (version 1.0). *PhysioNet* (2020).
RRID:SCR_007345. <https://doi.org/10.13026/c2294b>

Usability

- ▶ Evaluation: sample-based vs entire ICU stay evaluation
- ▶ Individual detections – alarm fatigue
- ▶ Develop scores:
 - PVA index
 - Alternatives: duration, effects, recovery?

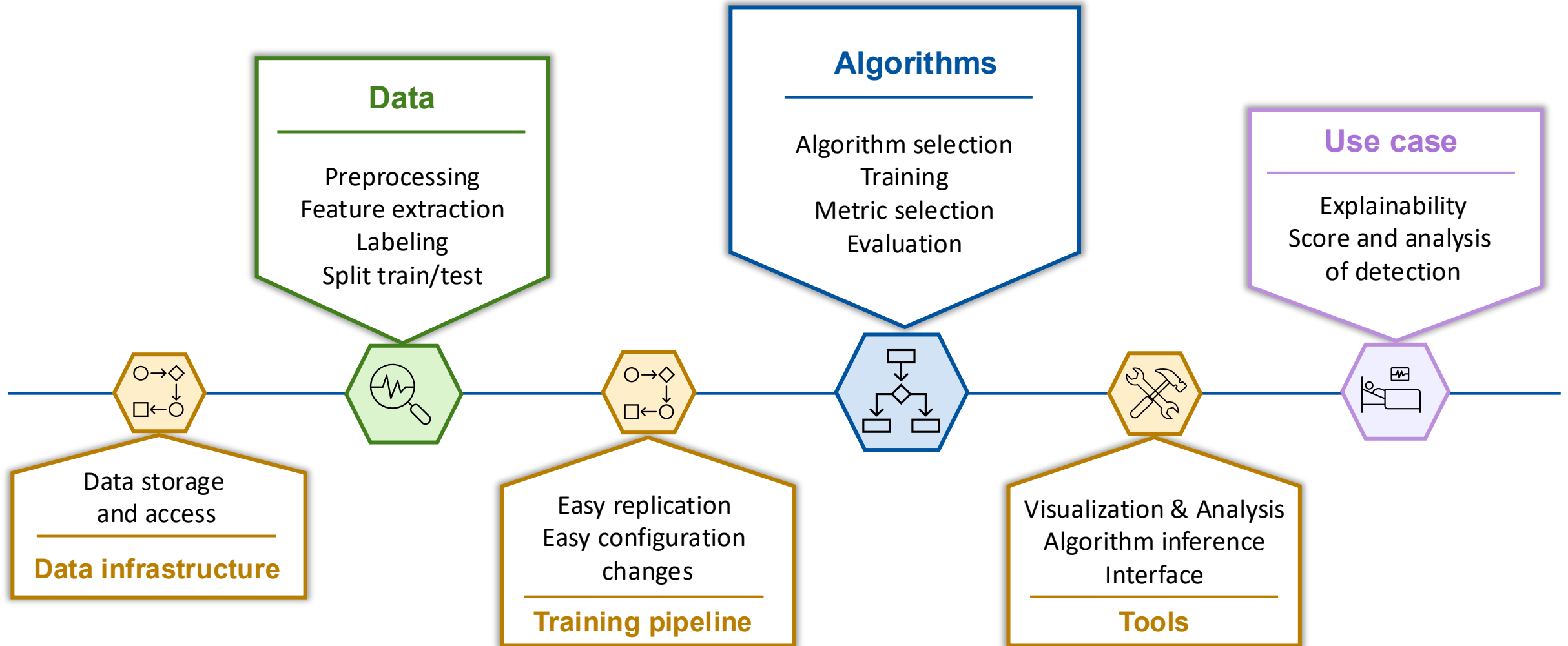
Explainability

- ▶ MDR requires some explainability
- ▶ Trust of different stakeholders? -> Many xAI papers, few evaluation with stakeholders
- ▶ Our own experience:
 - Good for development
 - More information provided: more questions asked

Conclusion



Overview



Research questions

