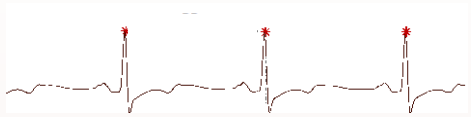
ReLearn: A Robust Machine Learning Framework in Presence of Missing Data for Multimodal Stress Detection from Physiological Signals

Arman Iranfar, Adriana Arza, David Atienza

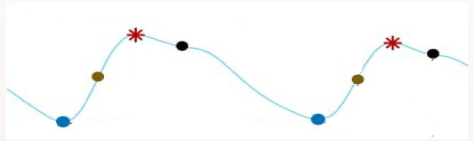
Embedded Systems Laboratory (ESL), Swiss Federal Institute of Technology Lausanne (EPFL)

Introduction

- Stress in moden societies
- Continuous and multimodal stress detection and recognition improves life quality [2]-[5]
  - Through wearable devices and embedded Machine Learning (ML) algorithms

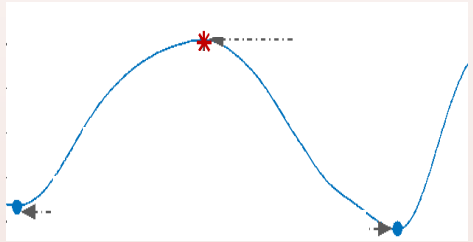


ECG (Electrocardiogram)



PPG (Photoplethysmography)

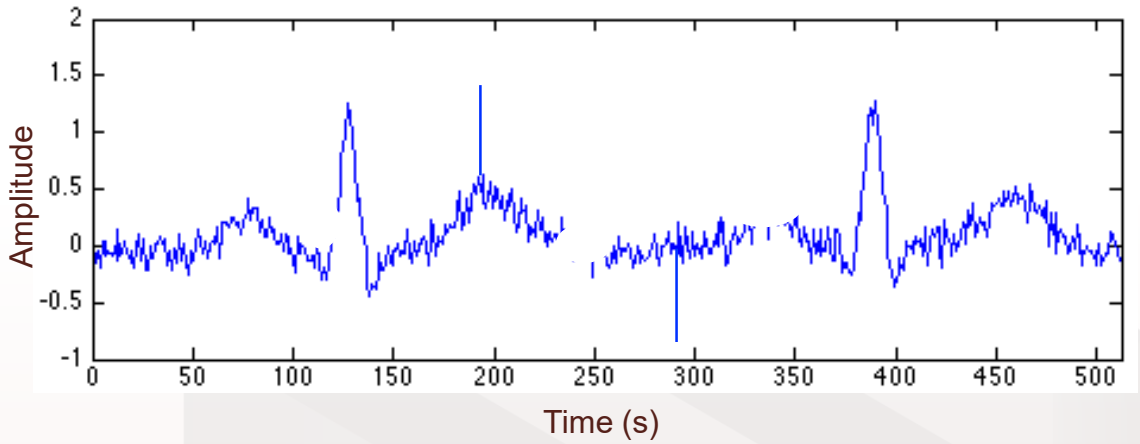
EDA (Electrodermal Activity)



RSP (Respiration)

Problem Definition and Motivation

- Common challenges of physiological signal acquisitions:
  - Noisy data
  - Missing data
  - Outliers



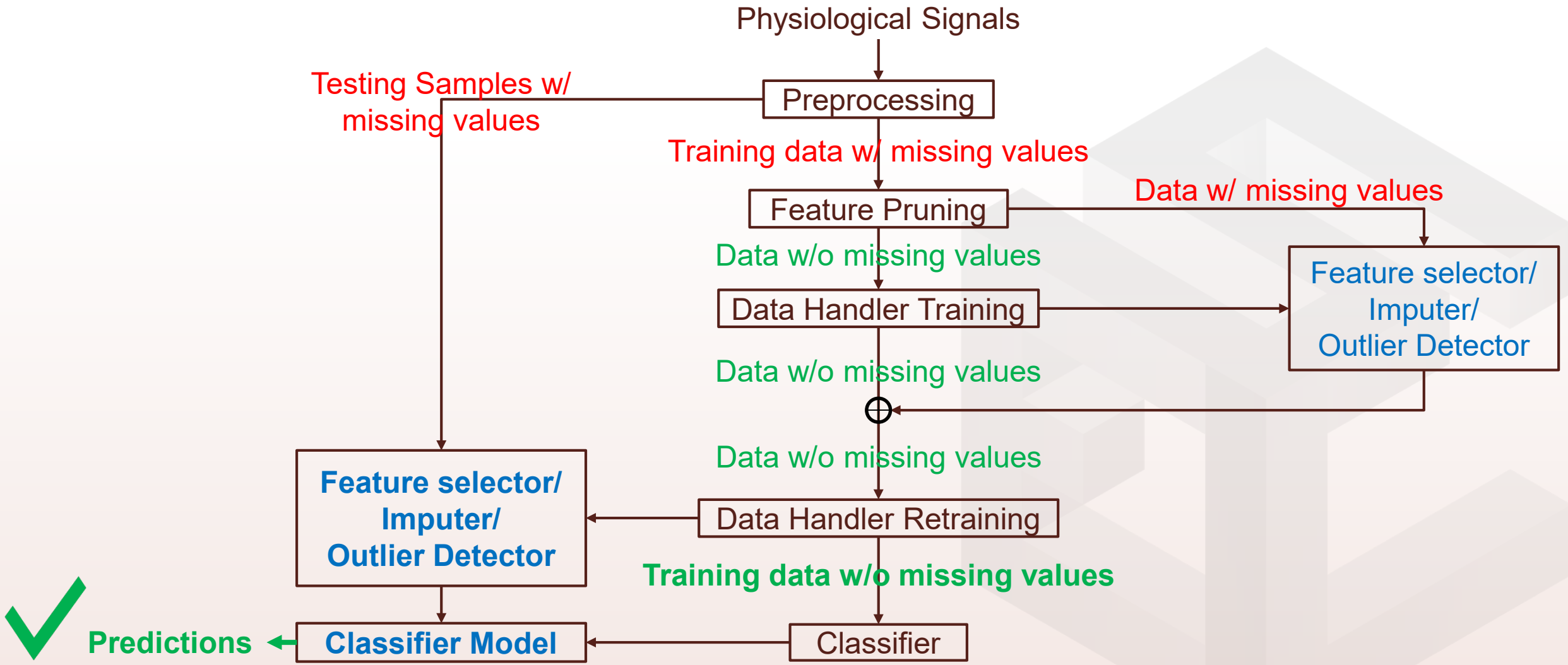
- Conventional remedies:
  - Filtering
  - Ignoring imperfect data and outliers
  - Data imputation
    - Last value
    - Mean/median value
    - Constant value

→
- No prediction for discarded segments
- Reduced prediction accuracy due to lack of sufficient data, esp. in Machine Learning solutions

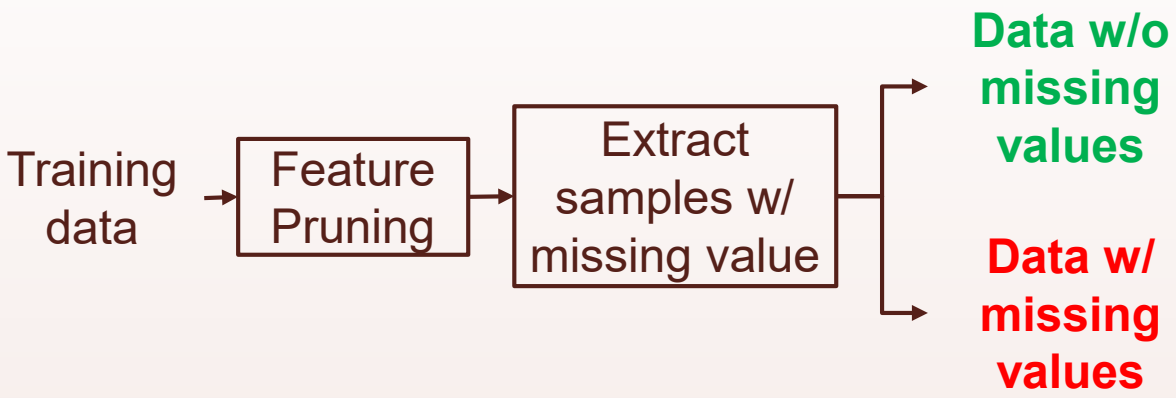
→
- Reduced prediction accuracy due to inaccurate data

A more comprehensive solution is requied to deal with outliers and missing data!

Proposed Framework: ReLearn

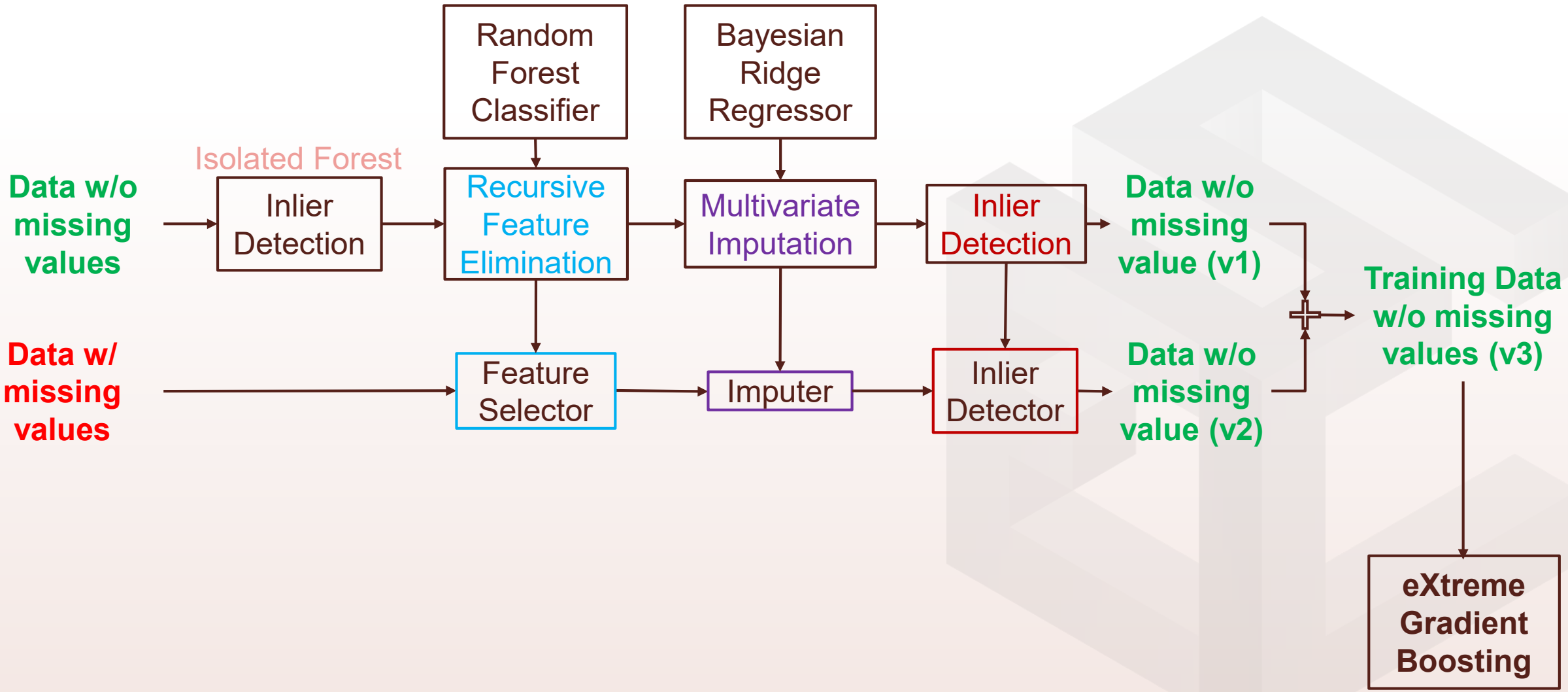


Data Preprocessing



- **Feature Pruning: remove a feature if more than x% of its samples are missing**
  - **Hyperparameter to tune**

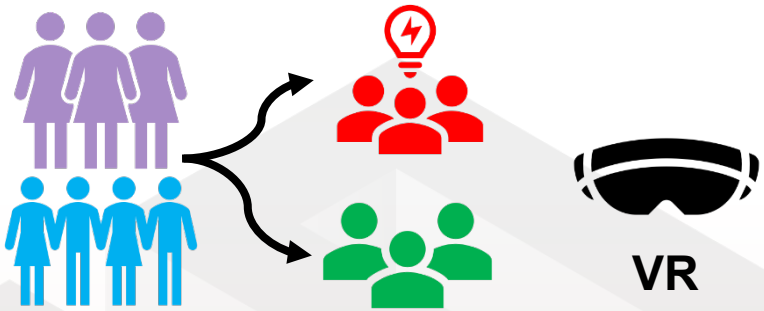
Data Handler and Main Classifier

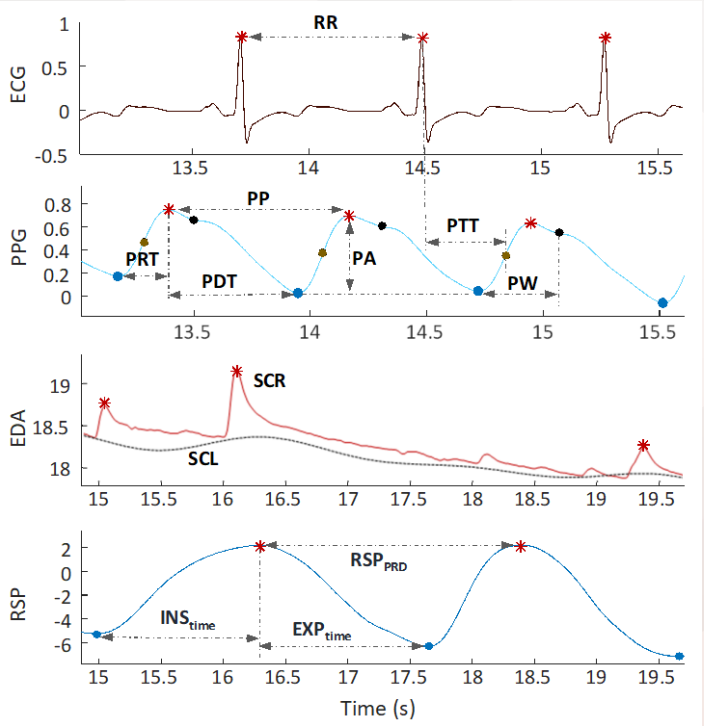


Experiment Protocol and Stress Database

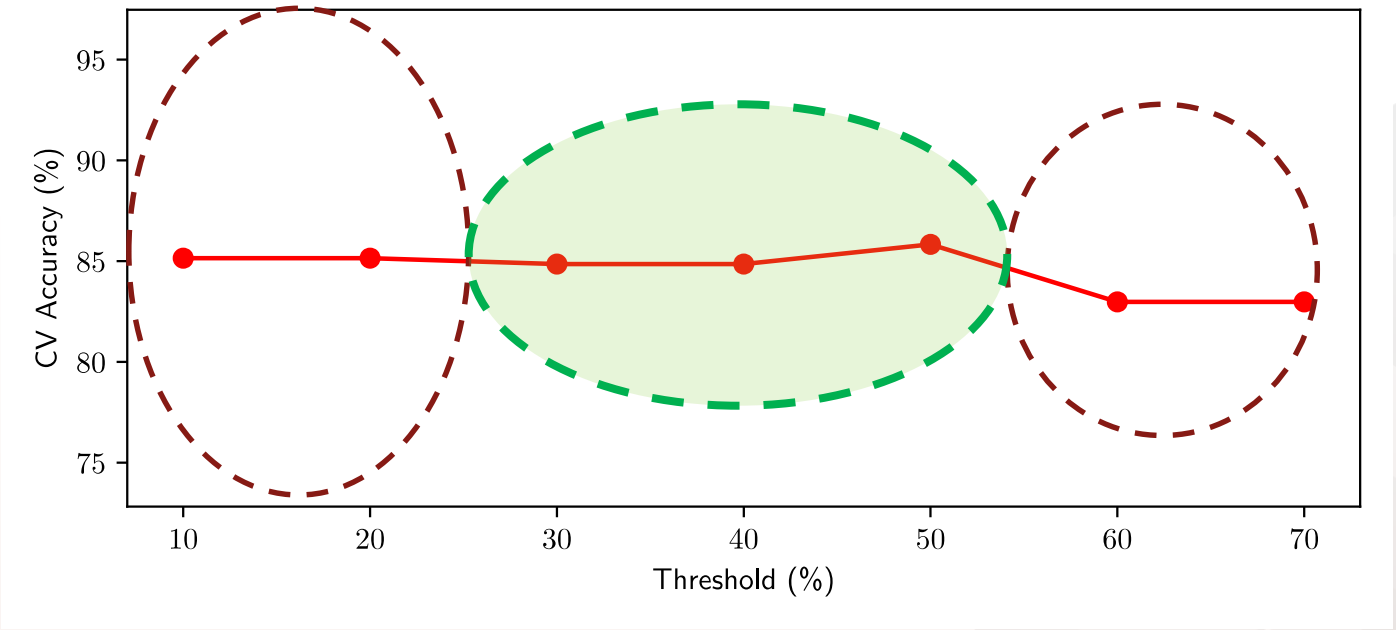
- Experimental Protocol:
  - 95 participant, male and female
  - Divided to two groups perfroming **stress** and **control** tasks

- Database:
  - ECG, PPG, EDA, RSP
  - 94 physiological features in the time and frequency domains extracted from time series in segmentation windows of 60s.





Results: Threshold Optimization



- With lower threshold, more features and samples are dropped: cleaner, but less data!
- With larger threshold, less features and samples are dropped: more unclean data!
- **Tradeoff required!**

Comparison with Conventional Methods

	CV (mean ± STD)	Inf. (All) (mean ± STD)	Inf. (Miss.)	#Feat
Drop NaN [13]	81.6 ± 12.0%	74.4 ± 30.0%	N/A	24
Mean Value [12]	53.9 ± 13.5%	50.2 ± 38.7%	51.1%	49
Last Value [12]	54.1 ± 12.4%	49.7 ± 38.2%	50.5%	36
ReLearn	**86.8 ± 6.4%**	**78.8 ± 25.4%**	**77.9%**	**38**

[13] Y. Bai, et al, “Fatigue assessment using ECG and actigraphy sensors,” International Symposium on Wearable Computers, 2020
[12] S. M. Jadhav, et al, “Artificial neural network based cardiac arrhythmia classification using ECG signal data,” International Conference on Electronics and Information Engineering, 2010.

Conclusion

- Machine learning algorithms require sufficiently large and clean data.
- Removing data with missing samples or using conventional imputation techniques lead to inaccurate prediction, especially for datasets with limited number of samples
- We proposed ReLearn, a machine learning framework for multimodal stress detection from physiological signals in presence of missing data and outliers
- Compared to conventional imputation techniques, ReLearn improved prediction accuracy by at least **26%** for missing samples.

Thank You!

For questions, please contact:
a.iranfar90@gmail.com

- **Acknowledgment:**

ML-Edge Swiss National Science Foundation (NSF) Research project (GA No. 200020182009/1)

DeepHealth H2020 Project (GA No. 825111)

ONR-G Award (Grant No. N62909-20-1-2063)