Frailty prediction using digital sensor data, an interpretable machine learning approach

Gaizka Pérez and Aleksandra Sjöström-Bujacz

Disclaimer

The analyses, their interpretation, and related information contained herein are made and provided subject to the assumptions, methodologies, caveats, and variables described in this report and are based on third party sources and data reasonably believed to be reliable. No warranty is made as to the completeness or accuracy of such third party sources or data.

As with any attempt to estimate future events, the forecasts, projections, conclusions, and other information included herein are subject to certain risks and uncertainties, and are not to be considered guarantees of any particular outcome.

Continuous monitoring with digital health technologies (DHTs) is a promising alternative enabling real-life assessments

- Frailty is defined as a clinical syndrome of increased vulnerability to stressors.
- It is measured with the **Edmonton Frail Scale (EFS):**
  - Adopted for use in countries throughout the world, primarily in research settings
  - 11-item clinician reported outcome
  - Good predictive validity (hospitalizations and mortality)

EFS: Edmonton Frail scale; DHTs: digital health technologies.
Rolfson, D. B., Majumdar, S. R., Tsuyuki, R. T., Tahir, A., & Rockwood, K. (2006). Validity and reliability of the Edmonton Frail Scale. Age and ageing, 35(5), 526-529.

Objectives

- The main objective of this work was to showcase how digital health technologies (DHTs) like electrocardiogram (**ECG**) sensors could be employed to derive important health measures such as **frailty**. Two different methodological approaches were taken:
  - **Traditional machine learning (ML) models** were tested after extracting tabular features from the ECG.
  - **Deep learning models** were tested after segmenting the raw time series in 4 second segments.
- The secondary objective was to balance model's predictive ability and interpretability by testing different models and post-hoc interpretability methods.

ECG: electrocardiogram; DHTs: digital health technologies; ML: machine learning.

Beyond the model's predictive ability, interpretability is crucial for stakeholders and regulatory bodies

Need for reasoning

➢It reaffirms that besides being accurate, ML model's predictions are trustworthy and accountable for their decisions.

Need for innovation

➢ML models may help us learn novel concepts and ideas (i. e. how ECG dynamics are related to frailty).

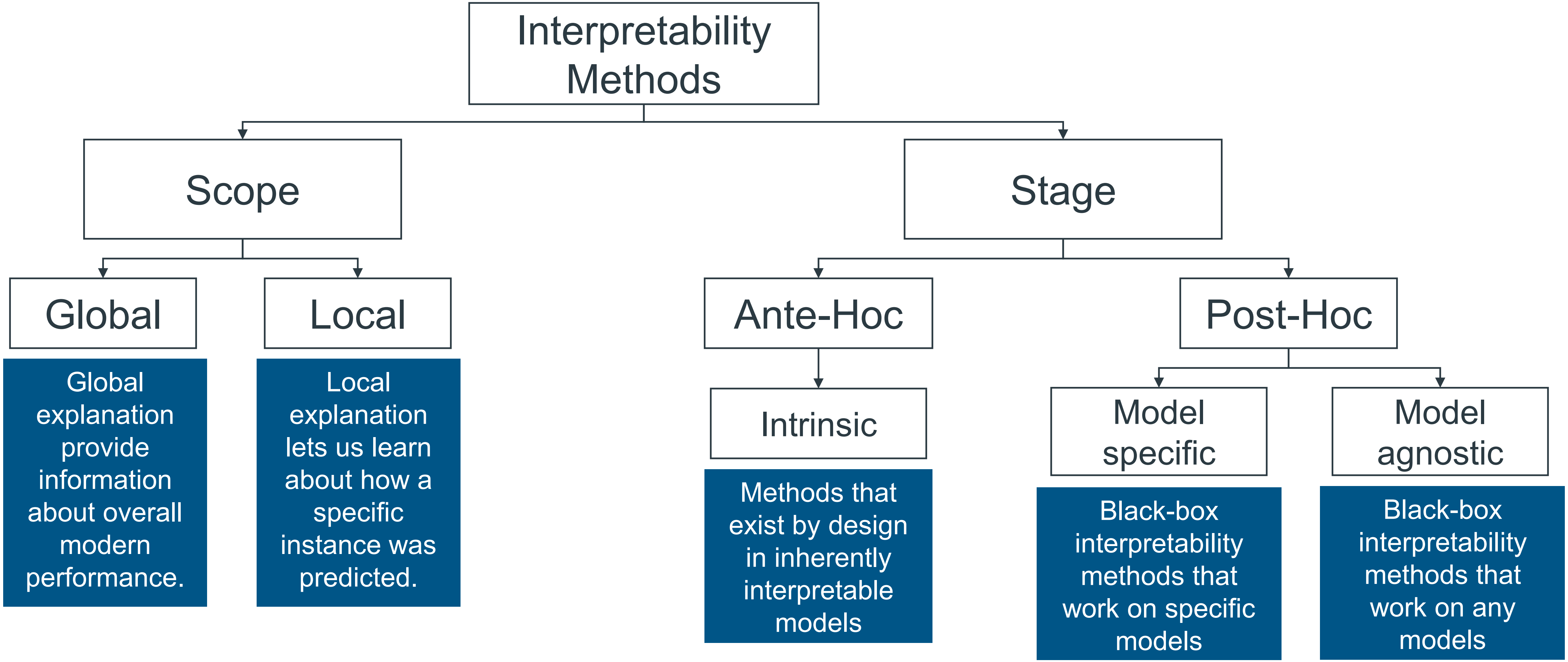
Need for regulation

➢Stakeholders and regulatory bodies demand model limitations and potential biases to be properly identified.

[…] because of the complex computational and statistical methodology […] understanding how AI models are developed and how they arrive at their conclusions may be difficult and necessitate methodological transparency **(FDA, 2025)**

ECG: electrocardiogram; ML: machine learning.
FDA (January 2025). Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products

Interpretability methods can be classified by Scope (broad vs. specific) and Stage (built-in vs. applied afterward)



Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... & Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. Cognitive Computation, 16(1), 45-74.

SHAP values are model agnostic post-hoc methods that allow us to interpret black-box models

- SHapley Additive exPlanations (SHAP) values are **model agnostic** and can provide both **global** and **local** explanations.
- SHAP assigns each feature an **importance value** for a particular prediction.
- They have some useful properties:
  - **Additivity:** ensures that the explanation fully accounts for the prediction
  - **Local accuracy:** allows us to perform local explanations
  - **Missingness:** robust to missing data
  - **Consistency:** increase monotonically based on marginal contribution

SHAP: SHapley Additive exPlanations.
Molnar, C. (2020). Interpretable machine learning. Lulu. com.

Publicly available data from the NCT04636970 study was used

Population:

- 80 patients on rehabilitation after open heart surgery

Protocol:

- **EFS score**
  - 0 to 5 was considered Non frail
  - 6 or greater was considered Frail
- **ECG** measured during **gait analysis**

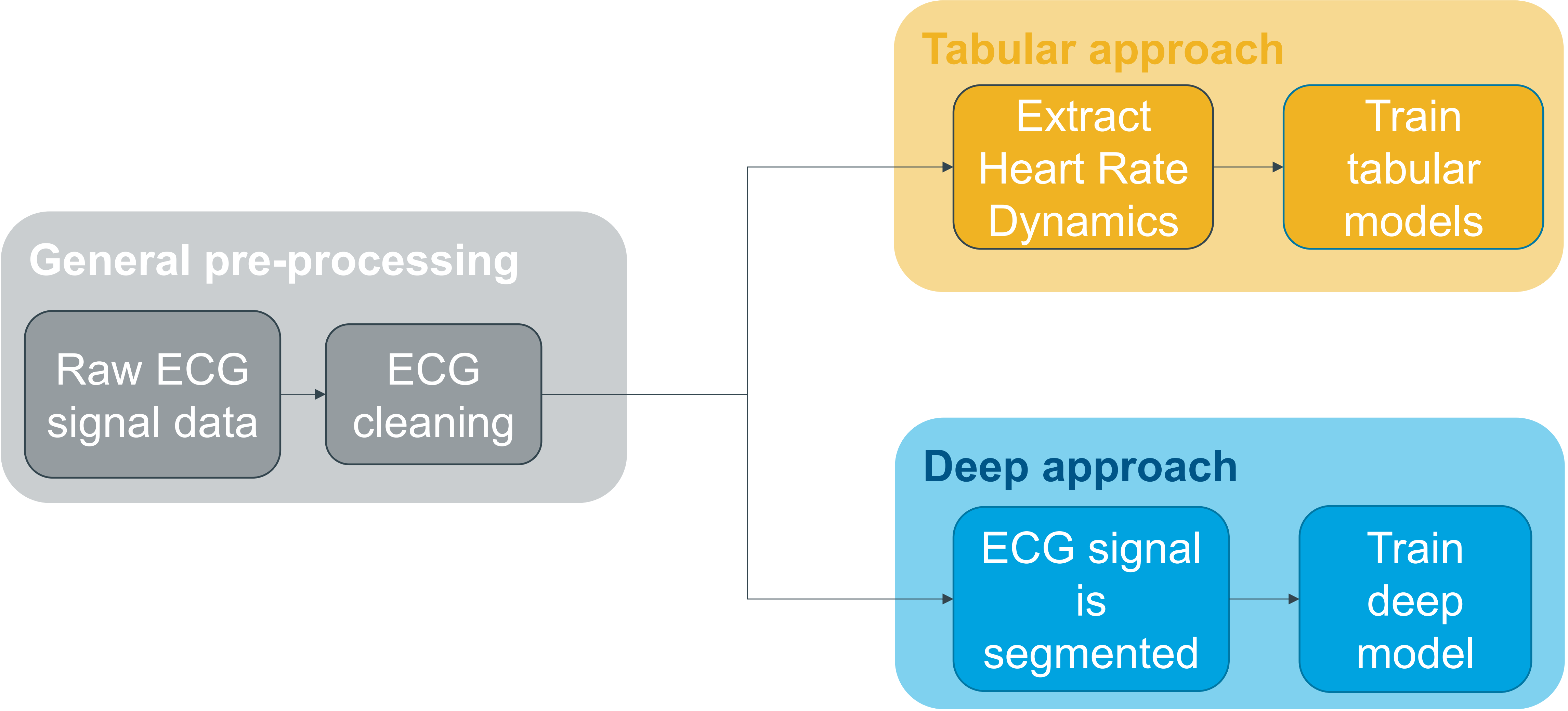
The ECG:

- ECG data was recorded using a **Polar H10**
- It was resampled to 130HZ

https://www.polar.com/uk-en/sensors/h10-heart-rate-sensor?srsltid=AfmBOoobcuR_E49boqOtUhvN593eIl4OxmumH4mNXPIuCMX5iG50jxPy

ECG: electrocardiogram; EFS: Edmonton Frail scale.
The data, with a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License was extracted from the following study: Sokas, D., Butkuvienė, M., Tamulevičiūtė-Prascienė, E., Beigienė, A., Kubilius, R., Petrėnas, A., & Paliakaitė, B. (2022). Wearable-based signals during physical exercises from patients with frailty after open-heart surgery (version 1.0.0). *PhysioNet*. https://doi.org/10.13026/mp8k-7p27.
The dataset was shared and available in PhysioNet: Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

Traditional machine learning models were compared to deep learning models



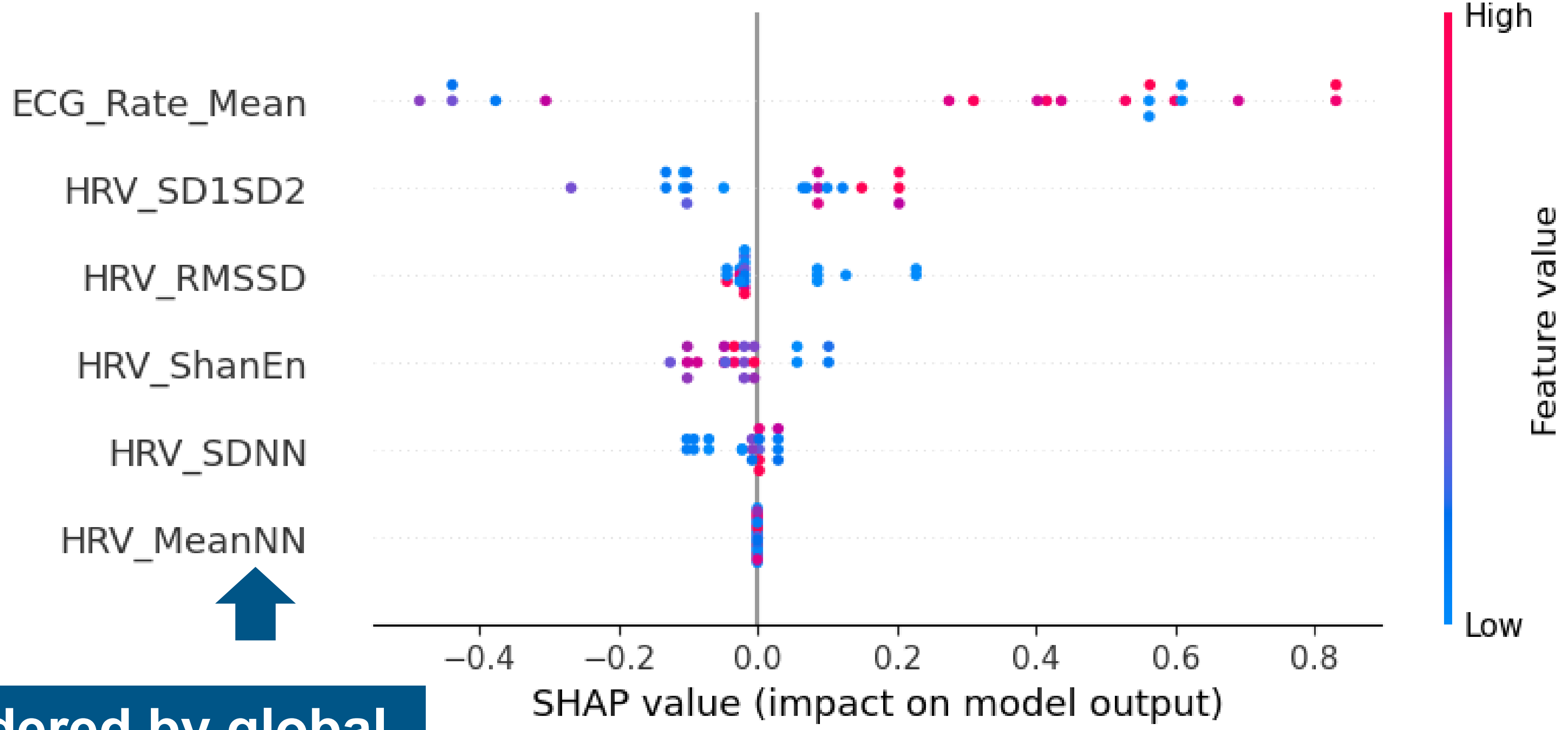
XGBoost showed the best results from all tabular models tested

Class	Models	Accuracy	Precision	Recall	F1
White Box models	Logistic regression	0.560	0.6	0.441	0.508
	KNeighbors	0.560	0.571	0.588	0.579
	DecisionTree	0.606	0.642	0.529	0.580
Black Box models	Random Forest	0.696	0.718	0.676	0.696
	XGBoost	0.757	0.764	0.764	0.764

TP: true positive; TN: true negative; FP: false positive; FN: false negative
Accuracy = (TP + TN) / (TP + TN + FP + FN); Precision = TP / (TP + FP); Recall = TP / (TP + FN); F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Mean heart rate was the most influential feature for predicting frailty

High absolute SHAP values indicate an influential value for the prediction



Color indicates feature magnitude

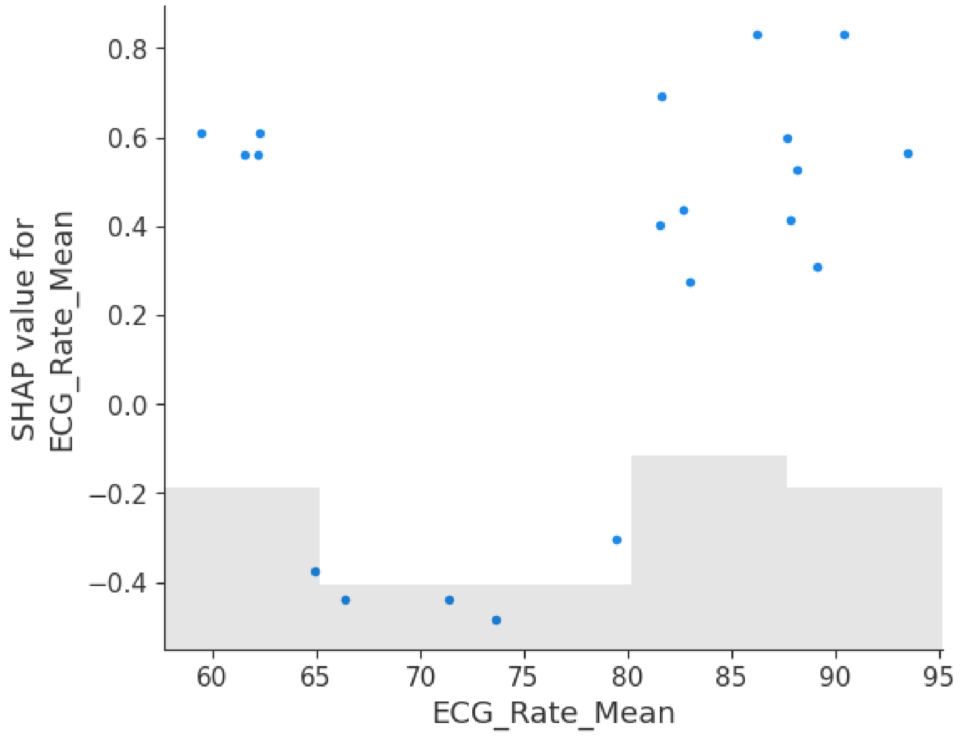
Ordered by global importance

SHAP: SHapley Additive exPlanations.

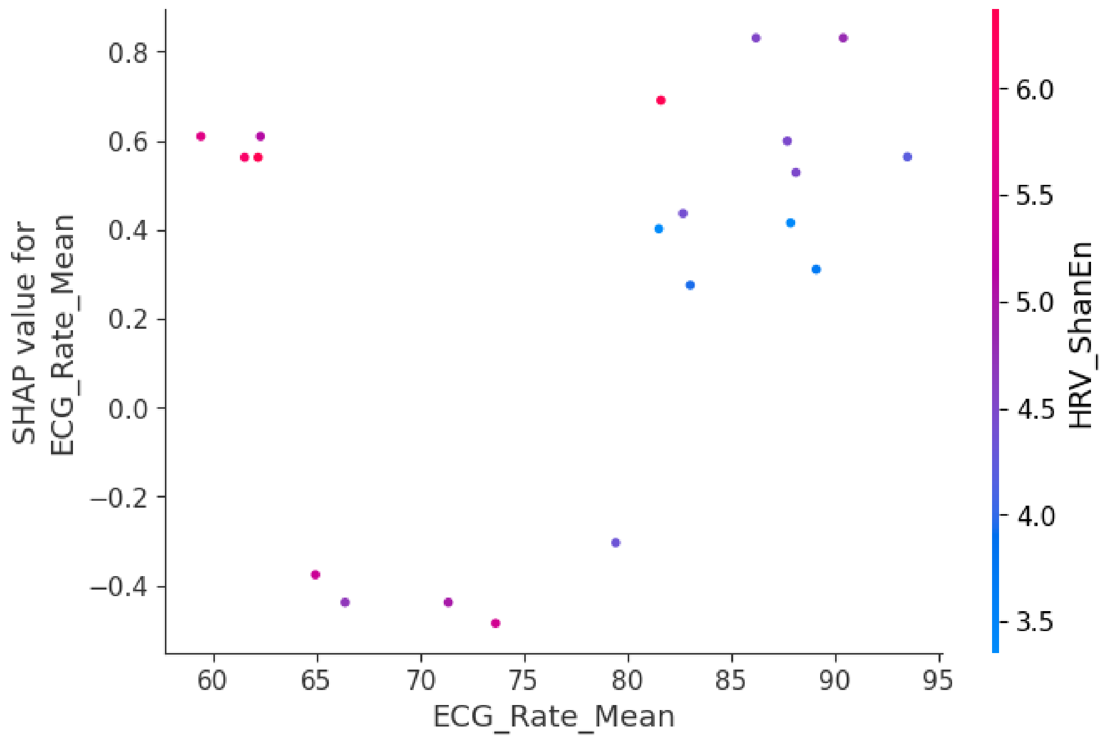
Mean heart rate appears to follow a quadratic relationship with frailty

Feature interactions can be analyzed by looking at the SHAP values

No interaction



Interaction



The deep learning model surpassed the performance of the best tabular model

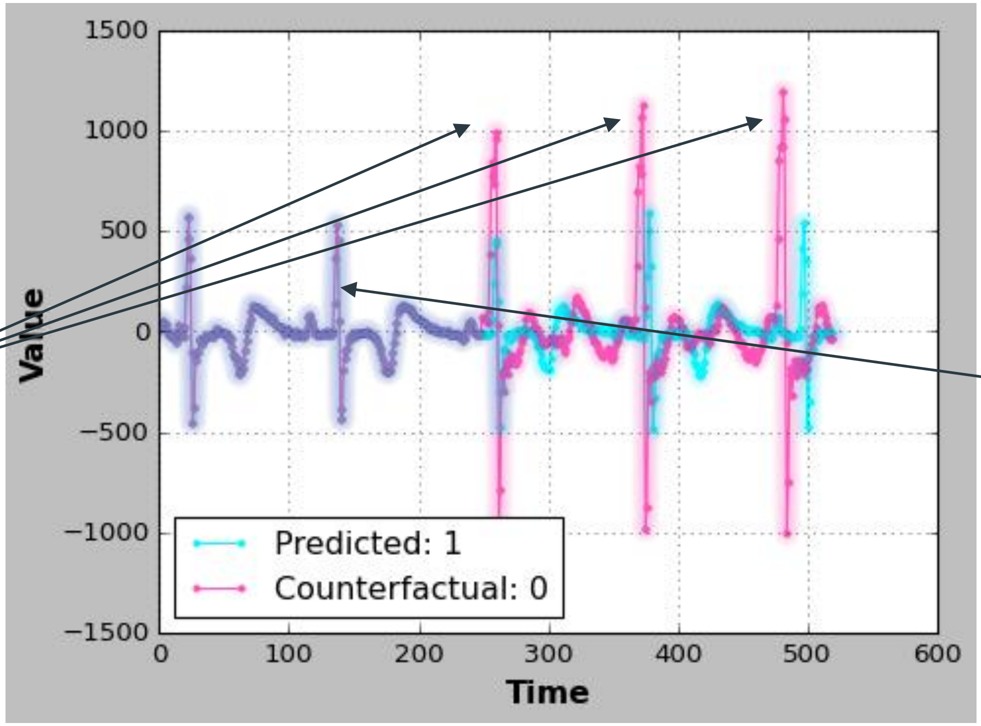
Class	Models	Accuracy	Precission	Recall	F1
Black Box models	Random Forest	0.696	0.718	0.676	0.696
	XGBoost	0.757	0.764	0.764	0.764
	1D CNN	0.903	0.881	0.864	0.850

1D CNN: one dimensional convolutional neural network; TP: true positive; TN: true negative; FP: false positive; FN: false negative
Accuracy = (TP + TN) / (TP + TN + FP + FN); Precision = TP / (TP + FP); Recall = TP / (TP + FN); F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Evolutionary Counterfactual Explanations for Time Series Classification (TSEvo) help us make counterfactual explanations

*A **non SHAP** based post-hoc local interpretation method for time series data*

In magenta we can see the changes in the time series that would make the model change its prediction to non frail



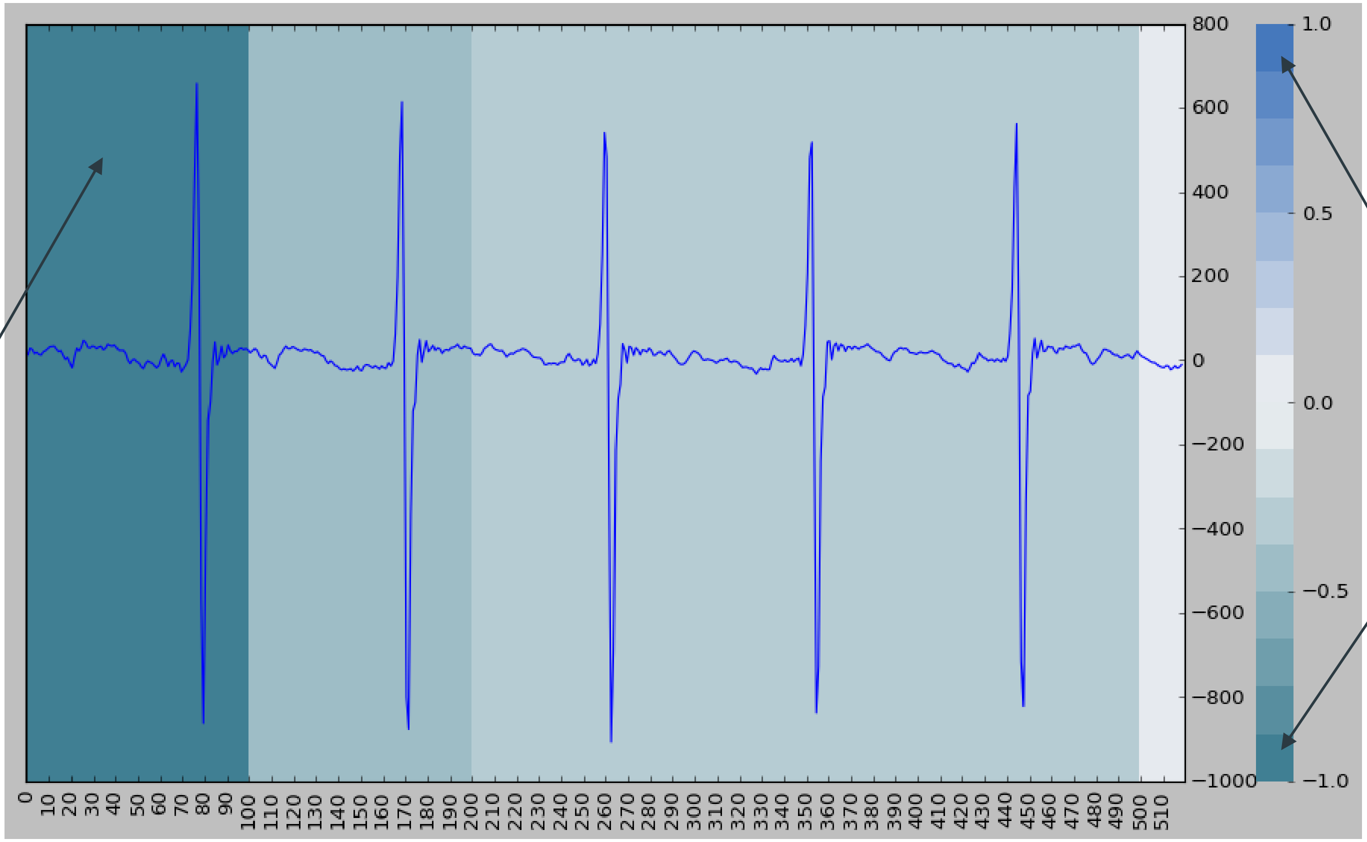
In purple/cyan we can see the original time series

TSEvo: Evolutionary Counterfactual Explanations for Time Series Classification.

Agnostic Local Explanation for Time Series Classification (LEFTIST) highlights which segments were more influential in the final prediction

*A **SHAP** based post-hoc local interpretation method for time series data*

More opaque colors represent segments that were more influential in the final prediction



Color hue indicates in which direction the segment was influential (Frailty – bright blue; Non frailty – dark blue)

LEFTIST: Agnostic Local Explanation for Time Series Classification.

Traditional ML models are more interpretable when tabular features can be extracted, while deep learning offers a good trade-off when understanding the time series itself is important

	Tabular models	Deep neural networks
Feature selection	They require previously defined features.	CNNs are able to autonomically detect features.
Data format	Require data in a tabular format.	Can work on tabular data, time series data, images or text.
Training time	Varies depending on model and data complexity.	With complex data formats can take days.
Model results	For tabular data, boosting algorithms such as XGBoost have been shown to perform best.	On tabular data, they are commonly surpassed by boosting algorithms.

ML: machine learning.

Conclusions

- Model selection is a key step, which influences both the prediction ability and the capacity to obtain insights from the model

- When dealing with tabular data intrinsically interpretable models may be preferred, when these don’t offer the performance needed **post-hoc interpretability techniques** can be employed

- When dealing with deep learning models post-hoc interpretability techniques are the only option for interpretability, which can help us get more granular local explanations