Interpretability for Decision Support in Intensive Care

MLCB

Intensive-Care Setting

Intensive-care clinicians are presented with large quantities of measurements from ***multiple monitoring systems***. The limited ability of humans to process complex information can hinder the ***timely recognition*** of patient deterioration. High numbers of rule-based alarms lead to ***alarm fatigue.***

Source: Nihon Kohden Inc.

Personalized Decision Support

Machine learning enables the development of ***personalized early-warning*** systems, harvesting ***all information*** available for an ICU patient (patient history, demographics, vital monitoring, lab results, machine-based patient support, treatment, -omics), and with a much ***lower false-alarm rate*** than conventional rule-based systems[1].

Interpretability in Clinical Practice

To be adopted in a ***clinical work-flow***, alarms issued must be interpretable. This is needed for the clinical personnel to (i) ***trust*** the alarm, and (ii) ***tailor the reaction*** to the alarm. Hence, there is a requirement to tune the trade-off between model interpretability and performance[2].

Interpretability for Timeseries

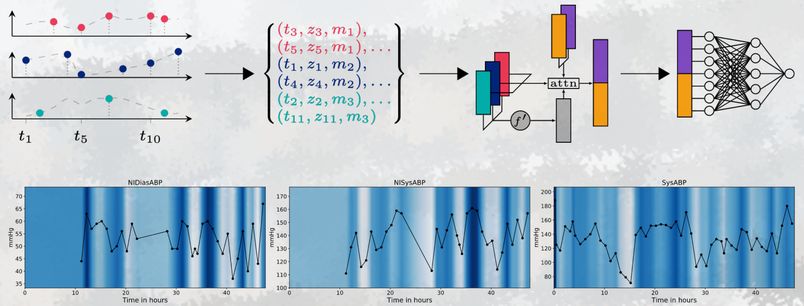
Medical timeseries pose ***multiple challenges*** for interpretability:

- ❑ ***Non-linearity:*** Models may not only react to a singular measurement, yet to a combination of measurements interacting across modalities and/or time
- ❑ ***Irregular sampling:*** How relevant points in time be determined, if not all (relevant) measurements are present at all times
- ❑ ***Non-random missingness:*** Both the presence and absence of a measurement may be relevant

Approaches

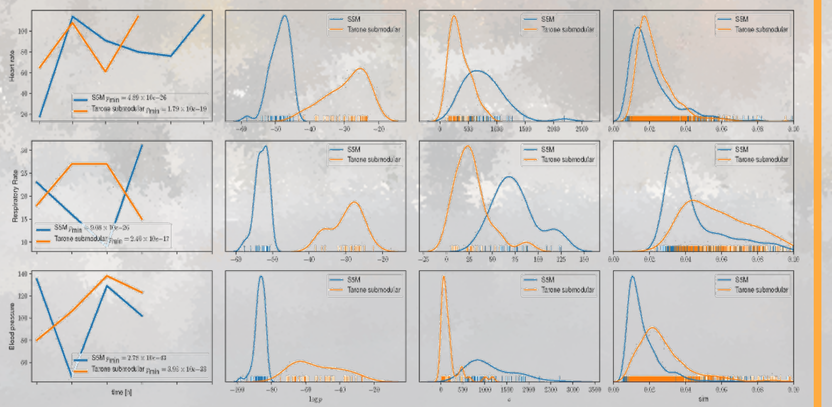
Attention Mechanisms

SeFT[3]: Input is a multivariate timeseries represented as ***sets of observations***. Each observation j is a tuple (t_j, z_j, m_j), comprising time t_j, value z_j, and moda-lity m_j. Set elements are ***summarized via a function*** f'. Conditional on both the summary and individual set elements an attention mechanism is applied to learn the importance of individual observations. Results of all ***attention heads*** are concatenated and used as input for the final classification layers.



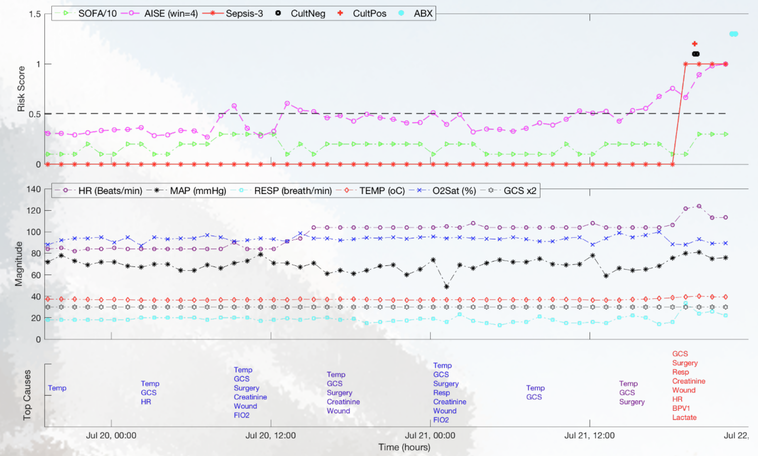
Shapelets

S5M[4]: Retrieves short ***subsequences*** that (i) occure in the data, (ii) are statistically ***significantly associated*** with a phenotype, and (iii) are of ***manageable*** quantity while maximizing ***structural diversity***. Diversity is achieved by pruning non-representative shapelets via submodular optimization. This increases the statistical power and utility.



Relevance Scores

RS[5]: Similar to ***saliency maps*** in deep learning, computed by taking the ***gradient of a risk score*** with respect to all input features and multiplied by the latter. The resulting relevance score deems an input feature relevant if it is both present and the model reacts to it. Relevance scores above the 95th percentile are reported.



Open Problems

	Non-linearity	Irregular Sampling	Non-random Missingness
SeFT		✓	
S5M	✓		
RS			

References

[1] Hyland, S. L., *et al.* "Early prediction of circulatory failure in the intensive care unit using machine learning." Nat. med. 26.3 (2020): 364-373

[2] Sendak, M. P., *et al.* "Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study." JMIR med. inf. 8.7 (2020): e15182.

[3] Horn, M., *et al.* "Set functions for time series." Int. conf. on machine learning. PMLR, 2020.

[4] Gumbsch, T., *et al.* "Enhancing statistical power in temporal biomarker discovery through representative shapelet mining." Bioinf. 36.2 (2020): i840-i848.

[5] Nemati, S., *et al.* "An interpretable machine learning model for accurate prediction of sepsis in the ICU." Crit. care med. 46.4 (2018): 547.

Interpretability in AI, *Banff International Research Station, Banff, CA, May 1st – 6th, 2022*

Swiss National Science Foundation

ETH zürich D BSSE