Building Interpretable Models for Time Series Classification

Hugues Turbé, Mina Bjelogrlic, Gianmarco Mengaldo, Medhi Namdar, Christian Lovis

MOTIVATIONS

Post-hoc Interpretability: Aim to identify features' contribution to model's outputs.
Interpretability methods attribute relevance to input features:

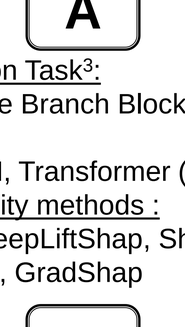
- Positive: input feature plays in favor of the model's prediction
- Negative: input feature plays against the model's prediction.

Interpretability is key for:

- **Building trust** in Artificial Intelligence (AI) based models[1]
- **Regulatory** aspect[2].

CHALLENGES

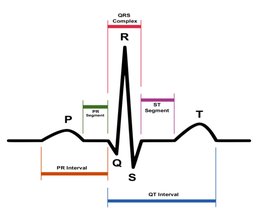
- Provide **meaningful insights** into the model



- **Evaluate** interpretability **methods**

CLASSIFICATION APPROACHES

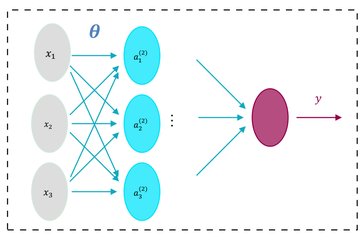
Insights into the model is dictated by model's inputs.



A

Raw signal

Pre-processing (e.g. Signal denoising)



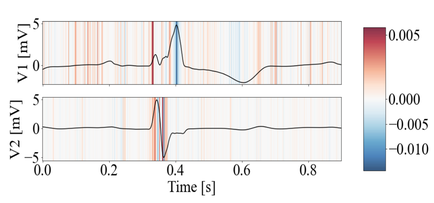


Figure 1: Relevance score produced using DeepLift along ECG signals.

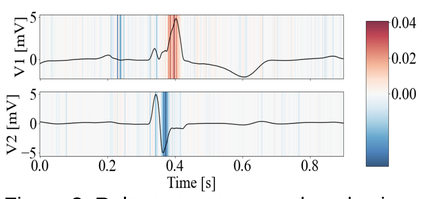


Figure 2: Relevance score produced using Shapley Value along ECG signals.

Pre-processing

Model training

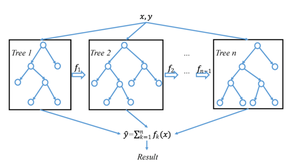
Post-hoc interpretability

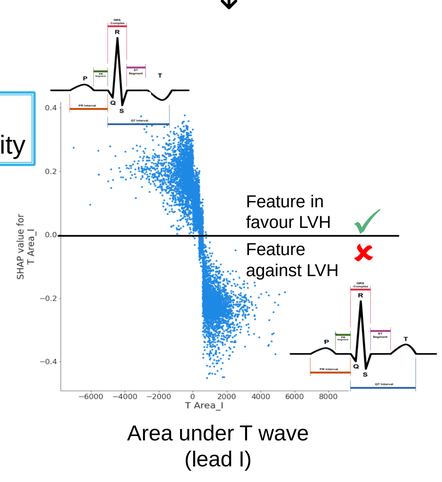
B

Extracted features

Expert-features extraction

	I	...	V6
R Amp	647	...	70
...	...	...	...
S dur	0	...	136





Area under T wave (lead I)

Figure 3: Relevance score vs. Area under T wave in lead 1.

METHODS

A

Classification Task[3]:
Right Bundle Branch Block (RBBB)
Models:
LSTM, CNN, Transformer (presented)
Interpretability methods :
DeepLift, DeepLiftShap, Shapley Value, KernelShap, GradShap

B

Classification Task[4] :
Left Ventricular Hypertrophy (LVH)
Model:
LightGBM
Interpretability methods:
Tree explainer

DISCUSSION

Different **interpretability methods** produce remarkably **different results**[3]:

→ **Evaluation** of **i**nterpretability methods is **critical.**

→ **Evaluation** should **assess** the **truthfulness** of the relevance **attribution**.

Relevance allows to compare features used by the model and cardiologists:
A: Shapley Value correctly identifies the R wave in lead V1 for LBBB classification.
B: Inversed T wave in lead I correctly identified as discriminator for LVH classification.
→ Build trust in models' predictions.

1. Shad, Rohan, et al. "Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging." Nature Machine Intelligence 3.11 (2021): 929-935.
2. European Commission, Directorate-General for Communications Networks, Content and Technology. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (2021).
3. Turbé, Hugues, et al. "InterpretTime: a new approach for the systematic evaluation of neural-network interpretability in time series classification." arXiv preprint arXiv:2202.05656 (2022).
4. Turbé, Hugues, et al. "A Lightweight and Interpretable Model to Classify Bundle Branch Blocks from ECG Signals." Challenges of Trustable AI and Added-Value on Health. IOS Press, 2022. 43-47.

SIMED@HUG SIMED@UNIGE

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

HUG Hôpitaux Universitaires Genève

SIMED