DIGITAL FINANCE

This project has received funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101119635

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

State Secretariat for Education, Research and Innovation SERI

Funded by the European Union

MARIE CURIE ACTIONS

DIGITAL

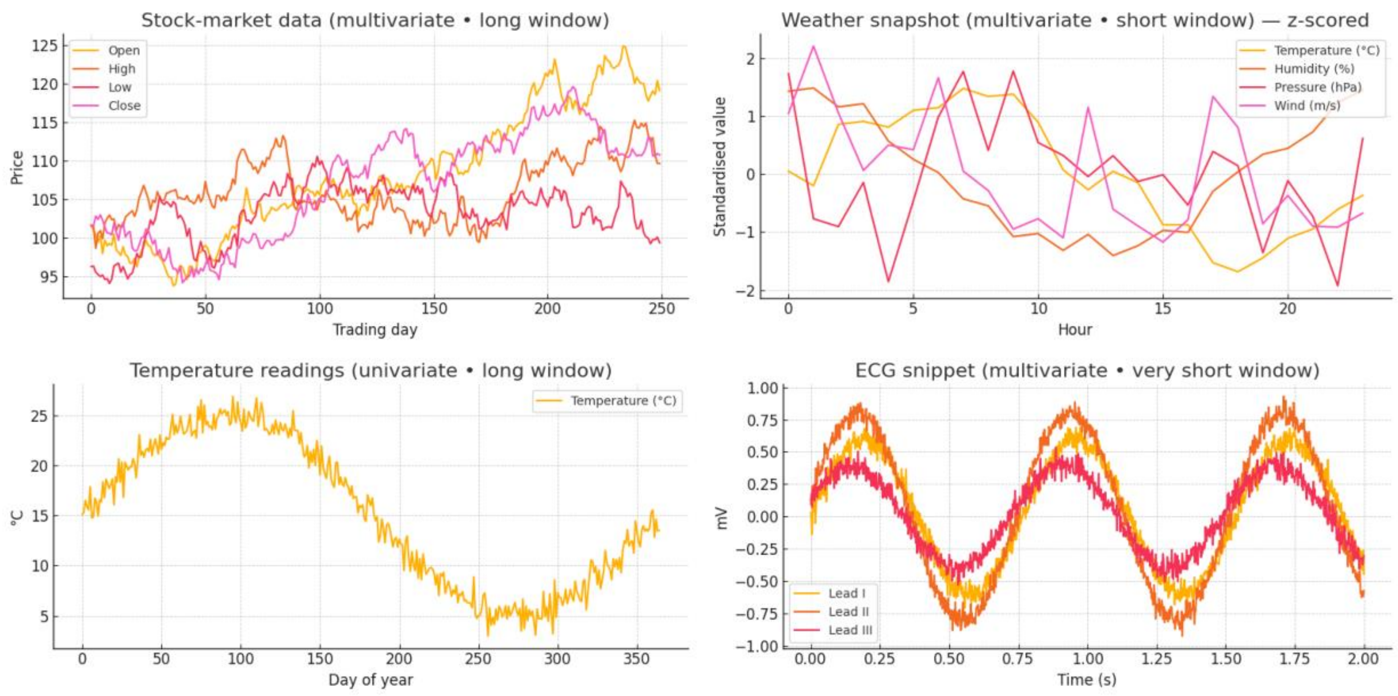
Time-Series xAI Methods

Faizan Ahmed

DIGITAL

Funded by
the European Union

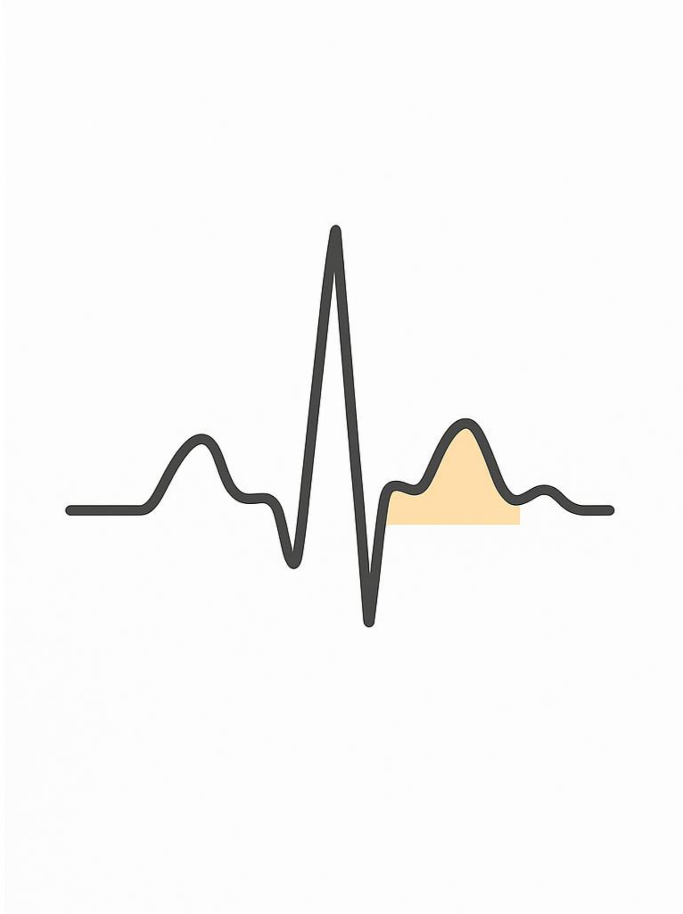
Time series



Number of observations collected over a successive period of time.

- Variable — *anything that changes over time*
- Time periods — *Can be daily, weekly, monthly, yearly*
- Variable Behaviour — *Quantifiable value*

Image vs TS



DIGITAL

Why Special methods for XAI?

Time series lack the spatial structure—no pixels, colours or shapes to **rely on.**

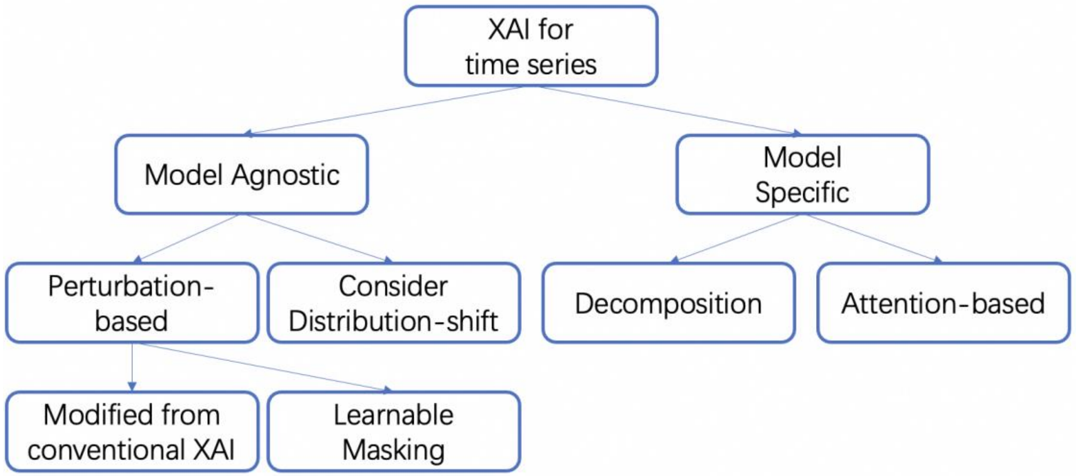
- Image saliency is visually intuitive; temporal signals rarely offer such anchors.
- Peaks and valleys alone seldom explain domain meaning; context over time is key.
- ECG example: subtle disturbances become clear only across several cardiac cycles.

Explanation methods for sequences must account for temporal dynamics.

Challenges

- Choosing an in-distribution baseline
- Interpretable temporal representation
- Capturing temporal interactions
- Managing computational cost

DIGITAL



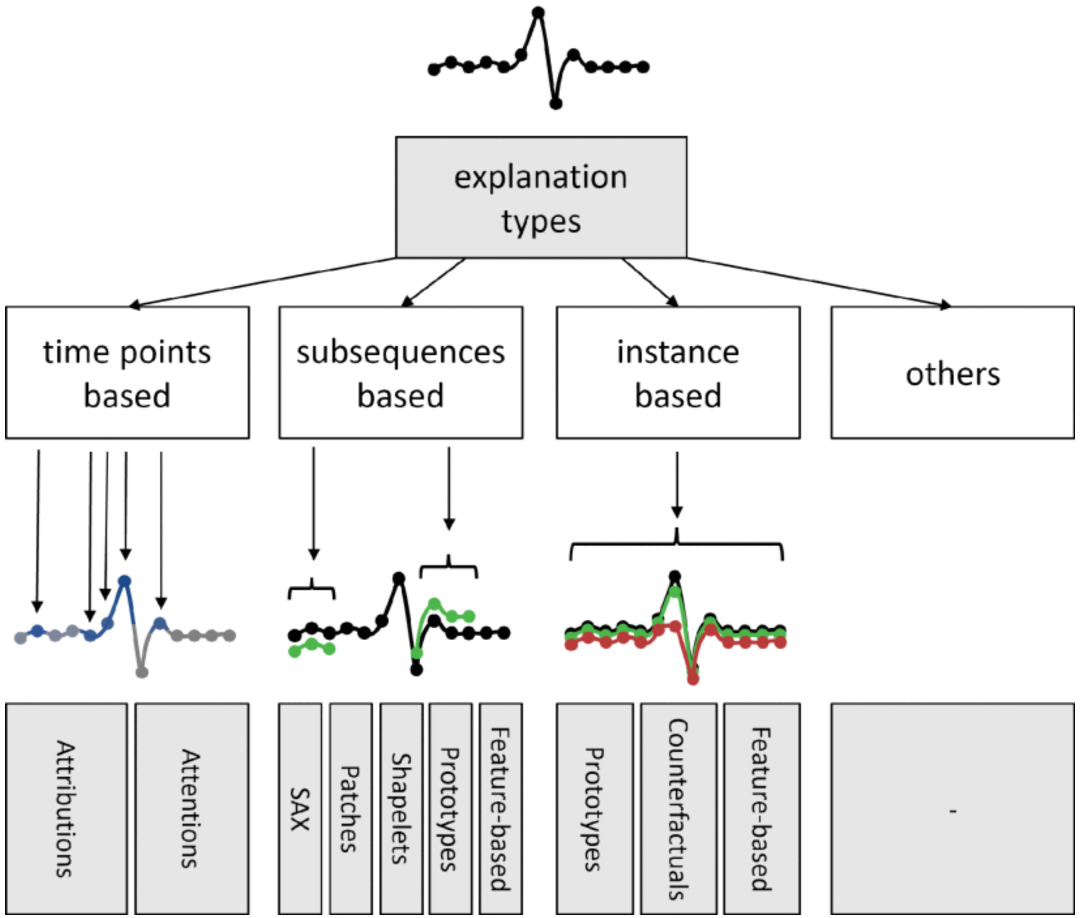
XAI for time series

TAXONOMY MODEL TYPE: GU, XINYUE & YANG, LINXIAO & SUN, LIANG. (2024). EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR TIME SERIES: A SYSTEMATIC SURVEY. 10.13140/RG.2.2.23062.56642.

DIGITAL

Family	Bucket	One-liner intuition
Model-agnostic	**Perturbation-based**	Mask or alter inputs and see how the prediction shifts (e.g., TimeSHAP, LIME-Segment, Dynamask).
	Distribution-aware	Replace inputs with *in-distribution* counterfactual samples; measure distributional shift (e.g., FIT).
Model-specific	**Decomposition-based**	Algebraically split the model into additive parts (e.g., Contextual Decomposition, ACD, REAT).
	Attention-score	Treat attention weights in RNNs/Transformers as importance indicators (e.g., RETAIN, TFT).

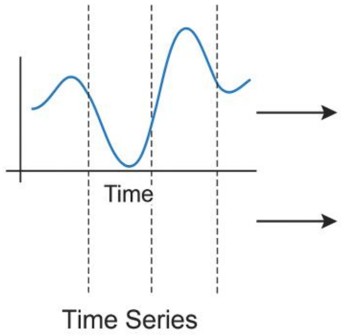
XAI for time series



- Taxonomy: Explanation Type
- See: https://ieeexplore.ieee.org/document/9895252

DIGITAL

XAI for TS: Windowing and Tabeling



Feature Engineering

Mean	Std. Dev.	Max
10.2	1.5	12,8
9.8	2.1	13,6
10.6	1.8	12,2
⋮	⋮	5.9
10.6	1.8	6,3
⋮	⋮	⋮

- **Window the raw signal** – slice it into fixed-length segments so each row in the future table corresponds to a "view" of the sequence.
- **Compute hand-crafted features** for each window
  - *Time-domain*: mean, variance, max/min, zero-crossings, slope, etc.
  - *Frequency-domain*: dominant frequency, band-power, spectral entropy, etc.
- **Result: a tabular matrix** (rows = windows, columns = features)
  - Now you can apply any tabular XAI tool (SHAP, LIME, feature permutation, global surrogate trees, etc.).
- **Pros & Cons**
  - **Pros:** interpretable features, fast inference, mature XAI support.
  - **Cons:** may discard fine-grained temporal patterns; requires domain knowledge to choose features.

DIGITAL

Time Series Grad-CAM

Algorithm 3: Gradient-weighted Class Activation Mapping

Input: (Multi/single variate) time series t, trained CNN, target class c

Output: Heat-map L_{GradCAM}

; /* Forward pass */

$A^k \leftarrow$ feature-maps of the *last conv* layer;

$S_c \leftarrow$ predicted score for class c;

; /* Backward pass */

$\frac{\partial S_c}{\partial A^k} \leftarrow$ gradients w.r.t. each map;

; /* Channel importance */

$\alpha_k = \frac{1}{T} \sum_t \frac{\partial S_c}{\partial A_t^k}$

* average only along the time axis

; /* Linear combination & ReLU */

$L_{\text{GradCAM}} = \text{ReLU}(\sum_k \alpha_k A^k)$;

; /* Upsample */

Resize L_{GradCAM} to the resolution of T and overlay;

Generate Heatmap?

Mandatory:

1. Assaf, R., & Schumann, A. (2019, August). Explainable deep neural networks for multivariate time series predictions. In *IJCAI* (pp. 6488-6490).

2. *J. Van Der Westhuizen and J. Lasenby. Techniques for visualizing lstms applied to electrocardiograms. arXiv preprint arXiv:1705.08153, 2017.*

DIGITAL

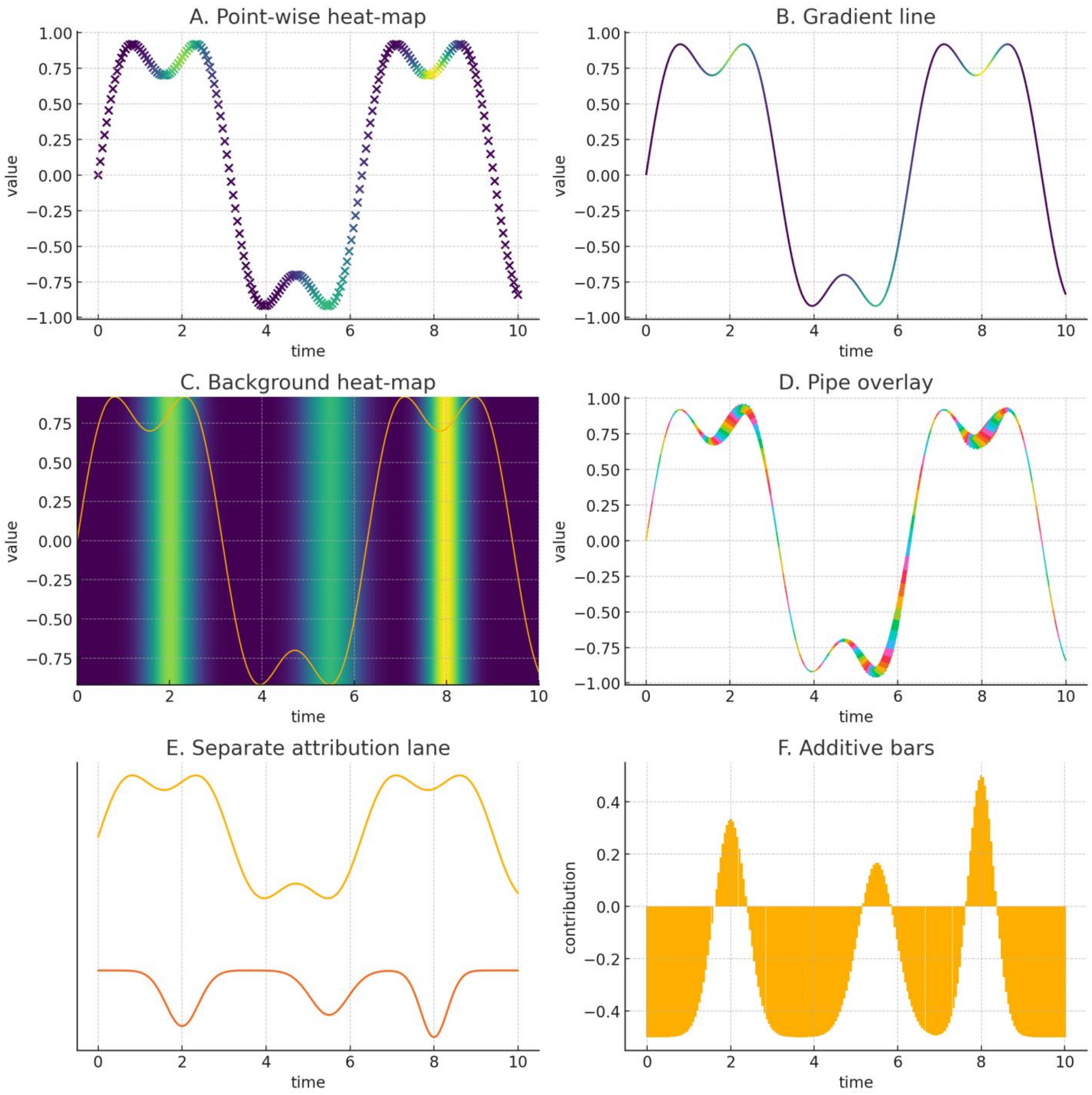
Family	Typical design	Strengths	Pain-points called out by the authors
A. Point-wise heat-maps	Put colour directly on the data marks (circles, dots).	extremely compact; preserves raw signal shape	heavy over-plotting, colour perception issues, later points over-draw earlier ones, hard beyond toy sequences
B. Gradient line segments	Encode relevance as a colour ramp along the poly-line.	keeps temporal ordering; no extra screen space	if the ramp is poorly chosen the signal itself becomes illegible; still juxtaposes high & low values without structure
C. Line-over-background heat-maps	Original curve in front, rectangular heat-map behind.	separates data & attribution layers; easy to add multiple attribution rows	overwhelms non-experts; small line vs huge heat-map; adjacent extremes visually clash
D. Dense pixel heat-maps (no signal)	Drop the line, show only a colour grid of relevance.	exposes recurrent patterns; good when the signal itself is distracting	eliminates temporal reference—experts struggle to relate colours back to real values
E. Pipe / tube overlays	Draw a variable-width, colour-coded "pipe" around the line.	width + colour jointly guide attention; low-relevance still visible	needs careful scale setting; may hide small fluctuations of the original series
F. Separate attribution lanes	Stack a second line plot (or small multiples) for relevance.	clean split—data intact, attribution legible; easy brushing & linking	relation between series & attribution requires eye jumps; less screen-efficient
G. Additive bar / arrow charts	Use SHAP additivity to draw positive/negative bars above & below the series, sometimes with arrows to compare models.	communicates direction (↑ helpful, ↓ harmful); supports multi-model comparison	only works for additive attributions; unsuitable for uni-variate series
H. Counter-factual-first workflows	Show "what would flip the prediction" examples first; drill down with attributions + what-if sliders.	aligns with Shneiderman mantra (overview → zoom → details); empirically easier for lay users	still a research vision; needs interactive tooling

DIGITAL

Visualizing TS explanation

Time Series Model Attribution Visualizations as Explanations

Visualizing TS explanation



DIGITAL

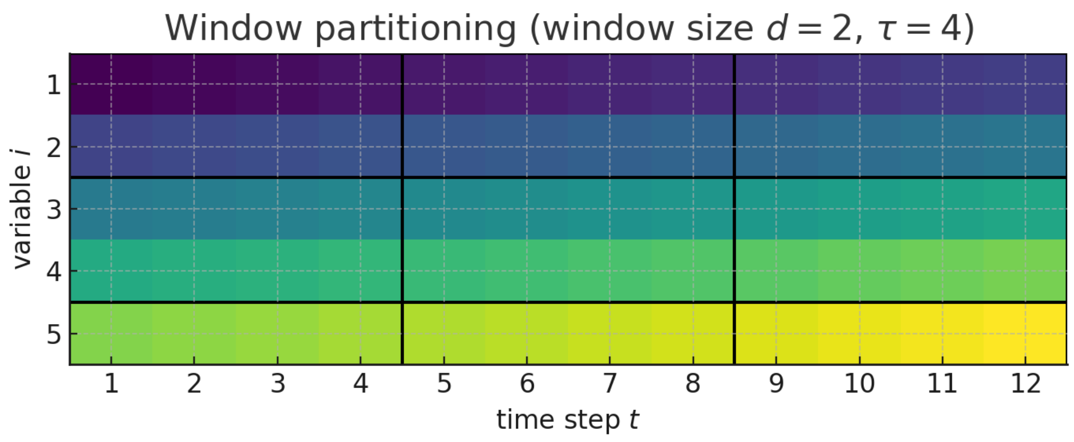
Explaining LSTM

TimeSHAP, WindowSHAp and C-SHAP

DIGITAL

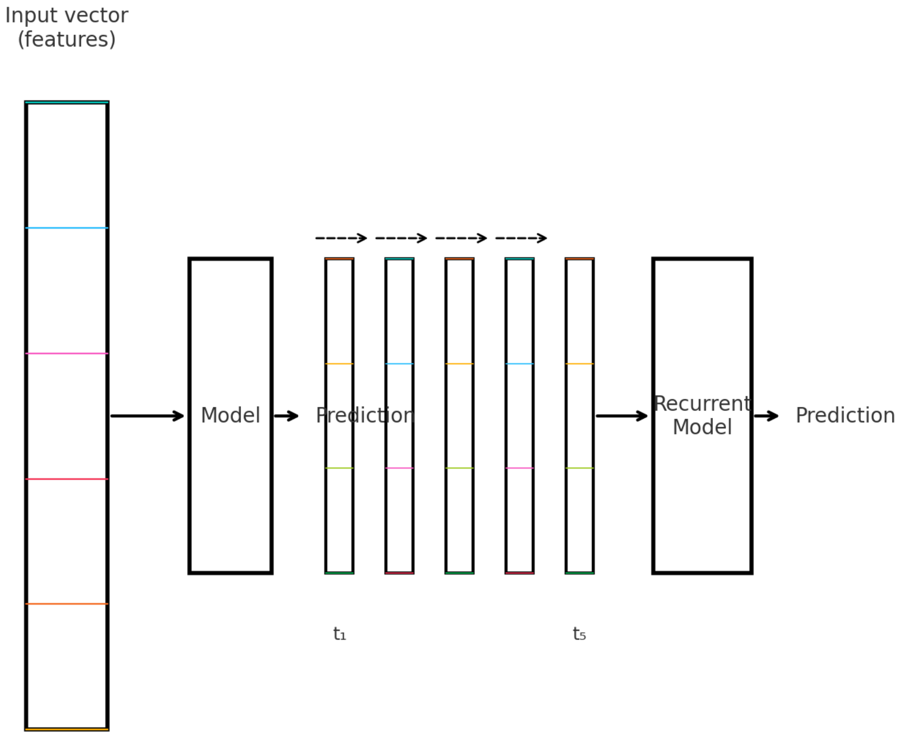
Shapley Values for Time Series

- Consider the multivariate time series $X \in \Re^{D\times L}$
- D is the number of variables
- L is the length of time series
- $\boldsymbol{\Delta} = \{(i,t) : 1 \leq i \leq D, 1 \leq t \leq T\}\}$ – set of all combination of time and variables.
- $\phi_{i,t} = \sum_{(S\subset\Delta\backslash\{(\mathrm{i,j})\}} \frac{|S|!(D\times L-|S|-1)!}{(D\times L)!} [v_{X^*}(S \cup \{(i,t)\}) - v_{X^*}(S)]$



DIGITAL

TimeSHAP



KernelSHAP on a static input vector
(one attribution vector)

TimeSHAP perturbs features × timesteps
(two-axis attributions)

- Tabular KernelSHAP treats the whole history as **one** feature vector → loses temporal context.
- RNNs, TCNs, Transformers output predictions **because of specific features at specific timesteps**.
- **Question:** "Which past events actually drove the prediction?"
- Requires attributions on **two axes** → *variables* × *timesteps*.

TimeSHAP

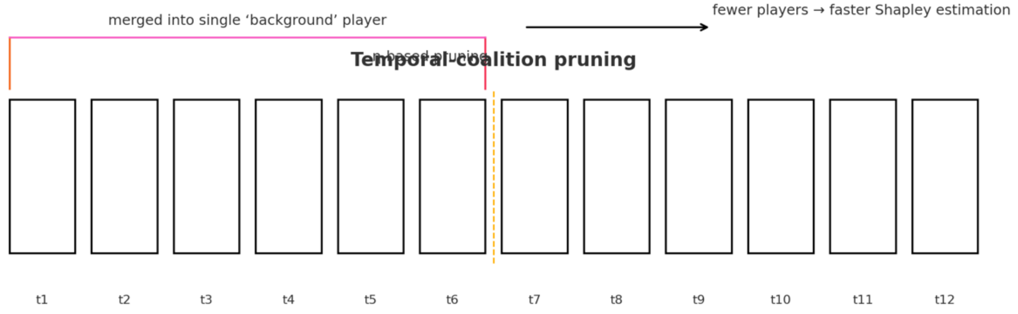
- Events are sorted by recency.

- Starting from the far past, merge events into one **background player** until

$$\sum_{j \in merged} \phi_{event_j} \leq \eta$$

where η is a user-set tolerance (e.g. 0.05).

- Reduces the exponential coalition space without violating Shapley axioms.



DIGITAL

TimeSHAP

Keep top-*k* rows (important features) and top-*k* columns (important timesteps).

Perturb only the **k^2 intersection cells** → quadratic, not exponential.

Produces fine-grained insights:

"This unusually large transfer amount at $t = k$ triggered the fraud alert."

DIGITAL

TimeSHAP

João Bento, Pedro Saleiro, André F. Cruz, Mário A.T. Figueiredo, and Pedro Bizarro. 2021. TimeSHAP: Explaining Recurrent Models through Sequence Perturbations. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21). Association for Computing Machinery, New York, NY, USA, 2565–2573. https://doi.org/10.1145/3447548.3467166

- **TimeSHAP** extends KernelSHAP to sequences, producing **feature-, event- and cell-level Shapley attributions** while remaining model-agnostic and post-hoc.
  - **Sequence-wide Shapley perturbations:** Extends KernelSHAP to *two* axes—features *and* timesteps—so you can ask "which past events and which variables actually drove the RNN's output?"
  - **Temporal-coalition pruning:** Groups the oldest, low-impact events into a single "background" coalition once their combined attribution falls below a tolerance η, slashing the exponential search space and runtime without losing Shapley guarantees.
  - **Cell-level zoom-in:** After isolating the few critical rows (features) and columns (events), it perturbs the individual cells at their intersections, yielding fine-grained attributions like "this unusually large transfer amount in event k triggered the fraud alert."

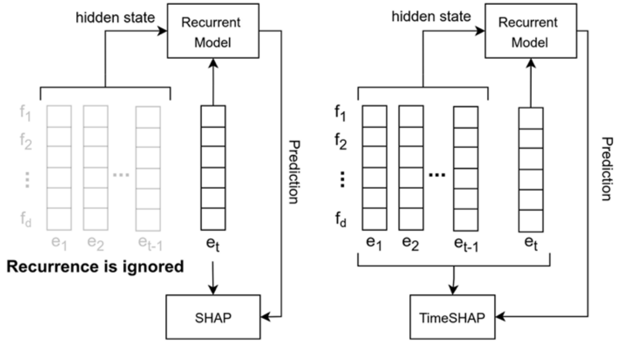
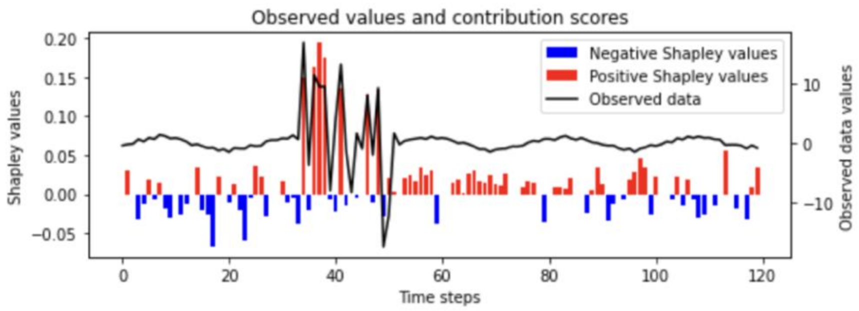


Figure 1: Current SHAP-based methods (left) only calculate attributions for a single input vector. TimeSHAP (right) applies perturbations throughout the input sequence.

DIGITAL

WindowSHAP

KernelSHAP

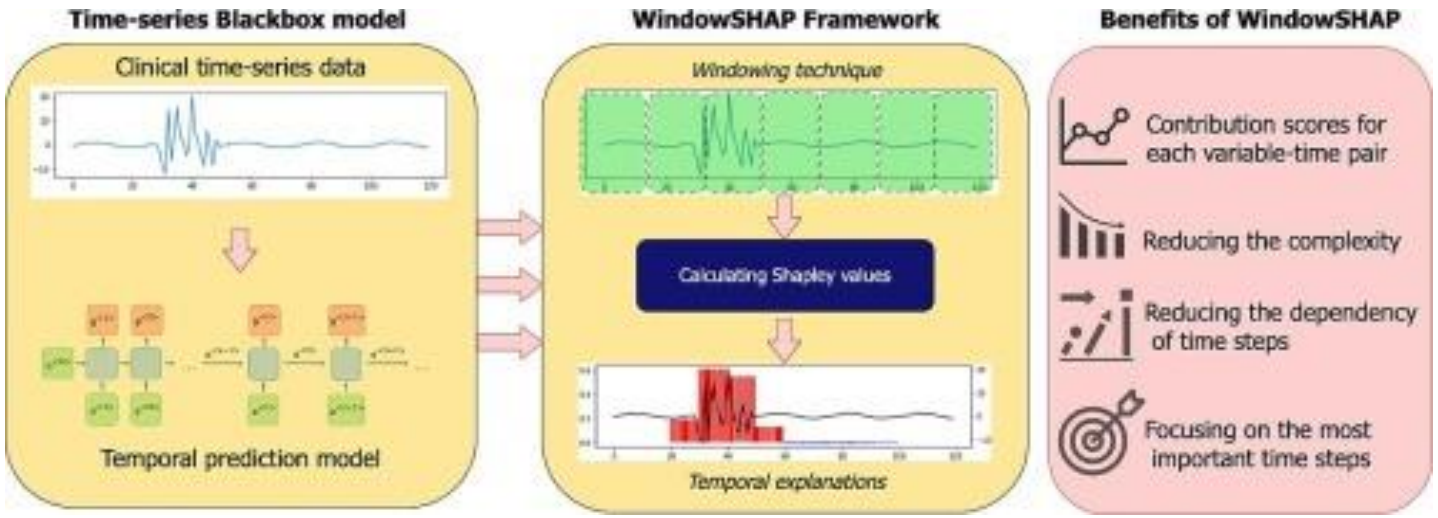


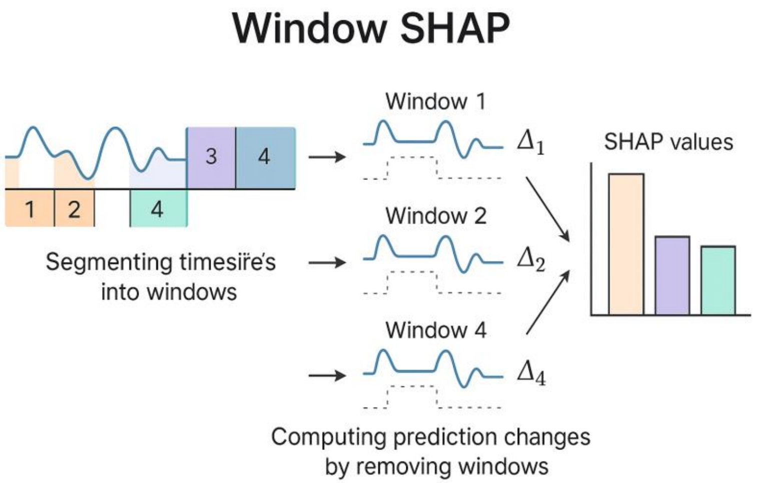
- Drawbacks (Kernel)SHAP for time series:
- Not originally intended to be used with time-series data.
- KernelSHAP approximates Shapley values and reduces computational time, but is still computationally expensive for high-dimensional data.
- Sequential data points are often highly dependent. For dependent features, their joint contribution is distributed among them, resulting in many small Shapley values.
    - Difficult to draw conclusions

DIGITAL

WindowSHAP

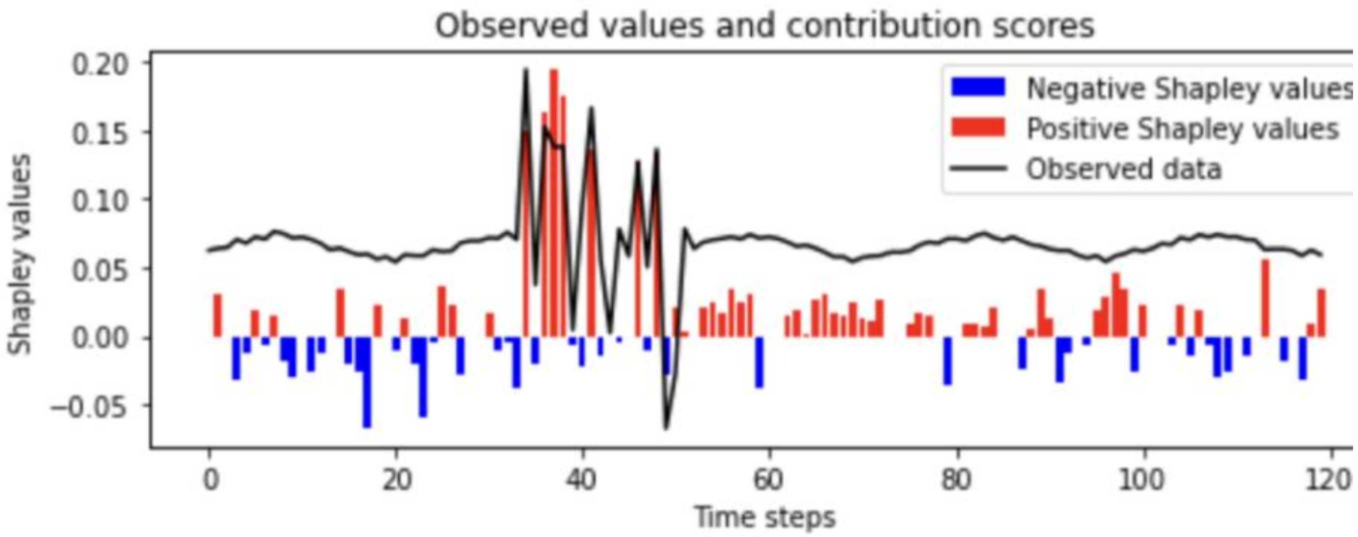
Nayebi, A., Tipirneni, S., Reddy, C. K., Foreman, B., & Subbian, V. (2023). WindowSHAP: An efficient framework for explaining time-series classifiers based on Shapley values. *Journal of biomedical informatics*, *144*, 104438.

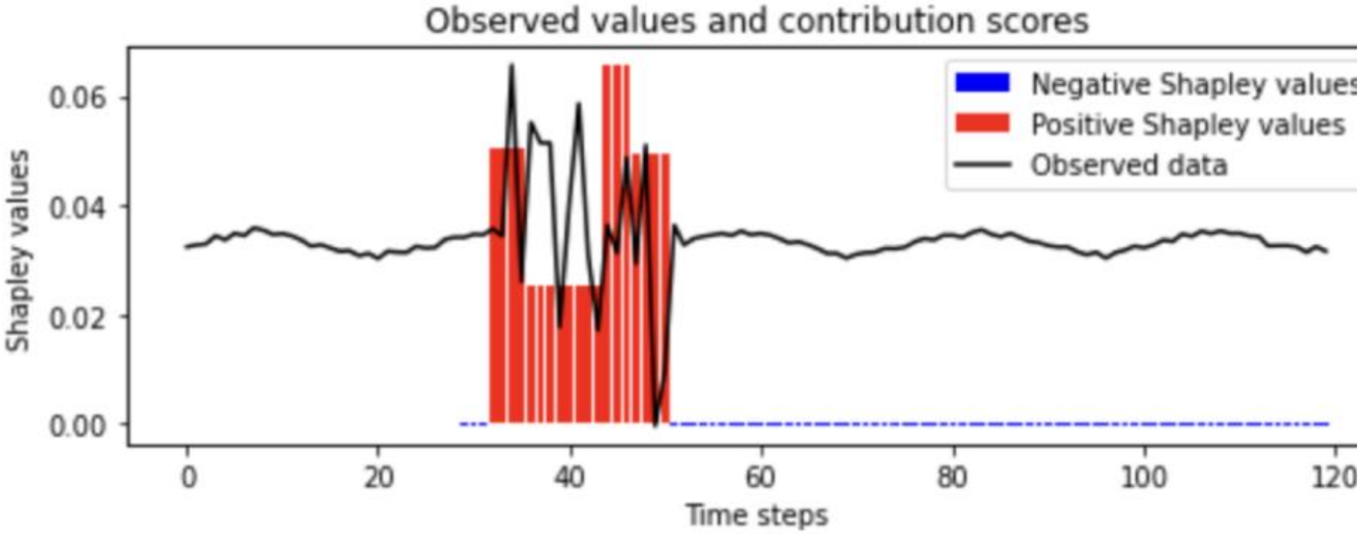




DIGITAL

WindowSHAP





DIGITAL

WindowSHAP

KernelSHAP:

- To mask a feature (=data point) replace it by an uninformative value
  - For example: zero, sampling from training data, …

WindowSHAP

- To mask a feature (=partition of data points) replace them all by an uninformative value
  - For example: zero, sampling from training data (subsequences), …

DIGITAL

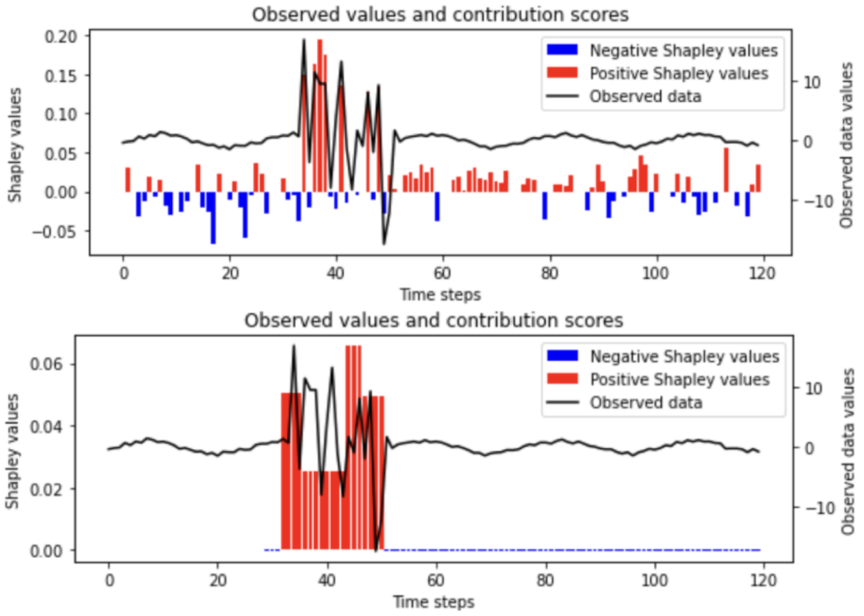
WindowSHAP

WindowSHAP solves these issues by partitioning data points and treating the partitions as features for SHAP:

- Partitioning means less features, so lower computational complexity
- Partitioning balances out small SHAP values and instead leads to more meaningful explanations

KernelSHAP

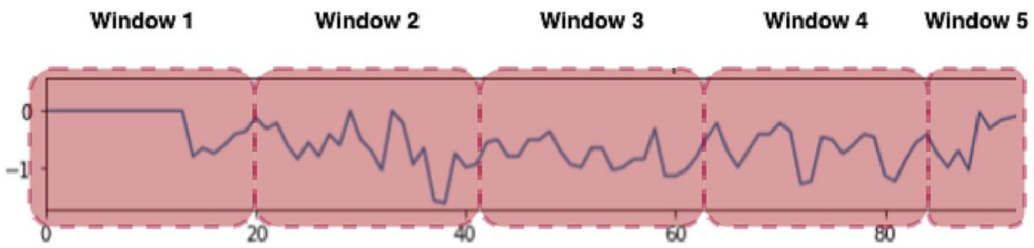
WindowSHAP



DIGITAL

WindowSHAP – partitioning

- Stationary WindowSHAP
  - Segment time series into adjacent fixed length windows



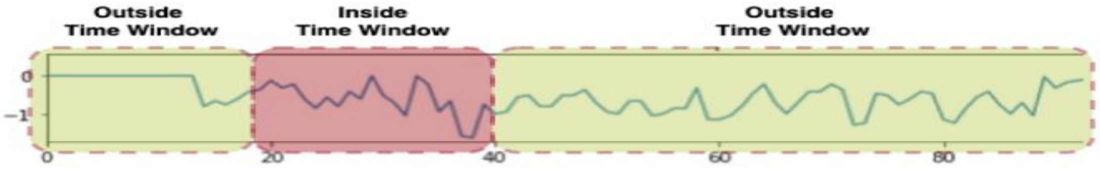
DIGITAL

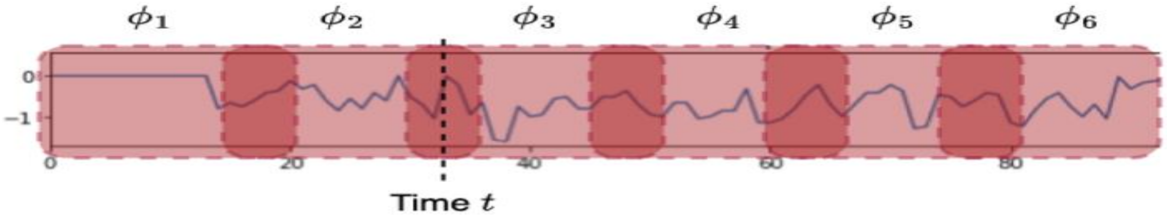
WindowSHAP – partitioning

Stationary WindowSHAP

Sliding WindowSHAP

- Segment time series into overlapping fixed length windows to mitigate boundary issues
- SHAP is repeatedly applied for each segment
- Average out SHAP values for overlapping windows

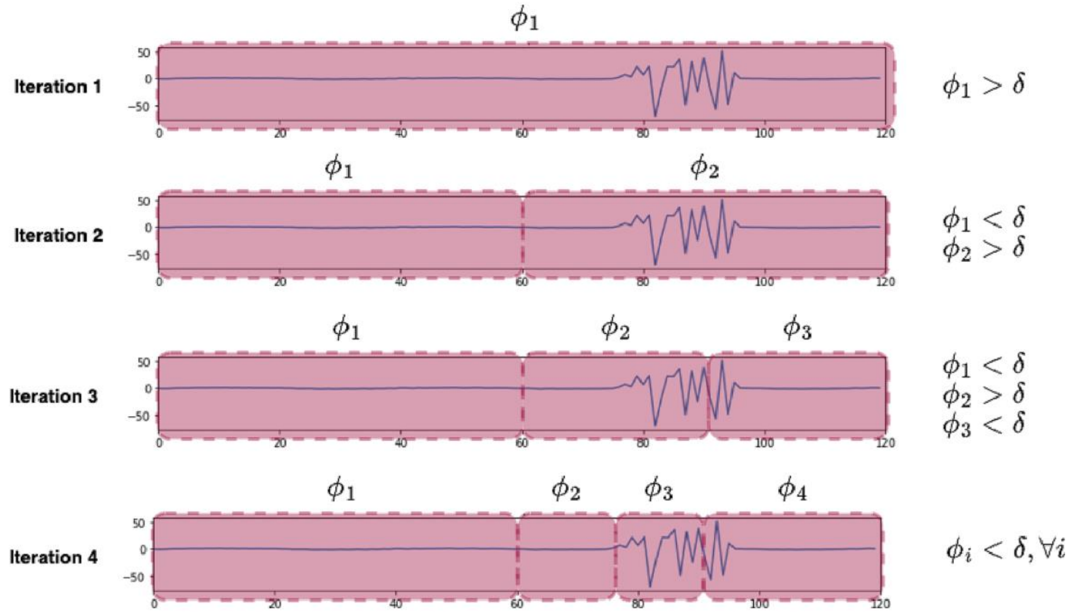




DIGITAL

WindowSHAP – partitioning

- Stationary WindowSHAP
- Sliding WindowSHAP
- Dynamic WindowSHAP
  - Flexible length windows
  - Repeatedly apply WindowSHAP and split partitions with high SHAP values



DIGITAL

WindowSHAP

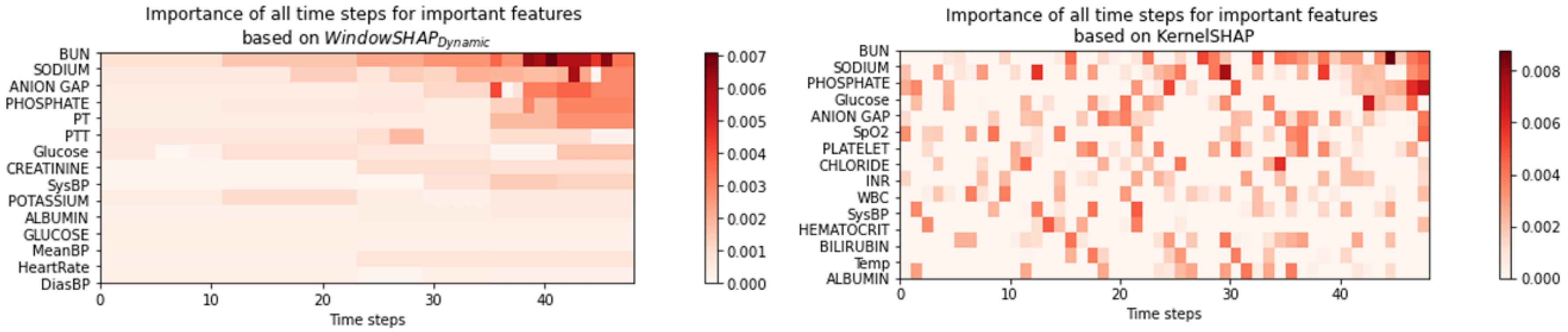
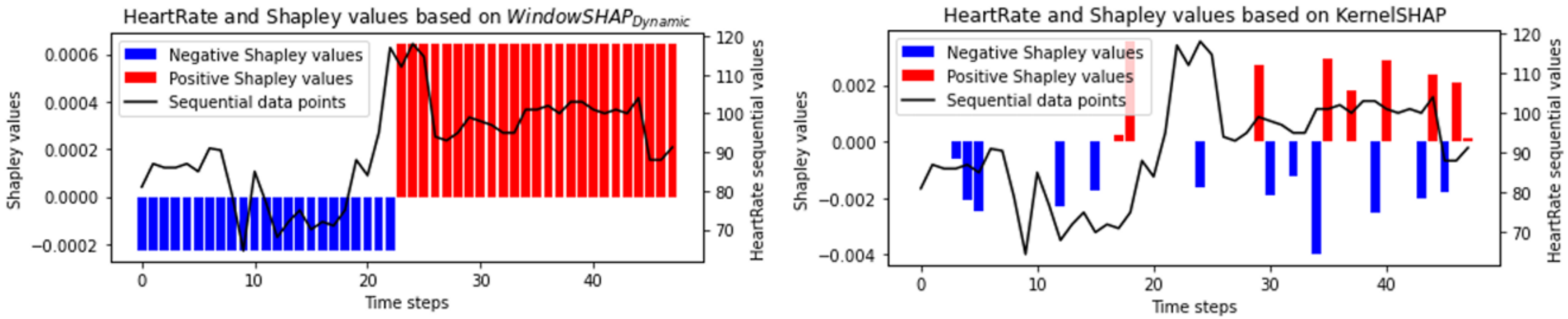
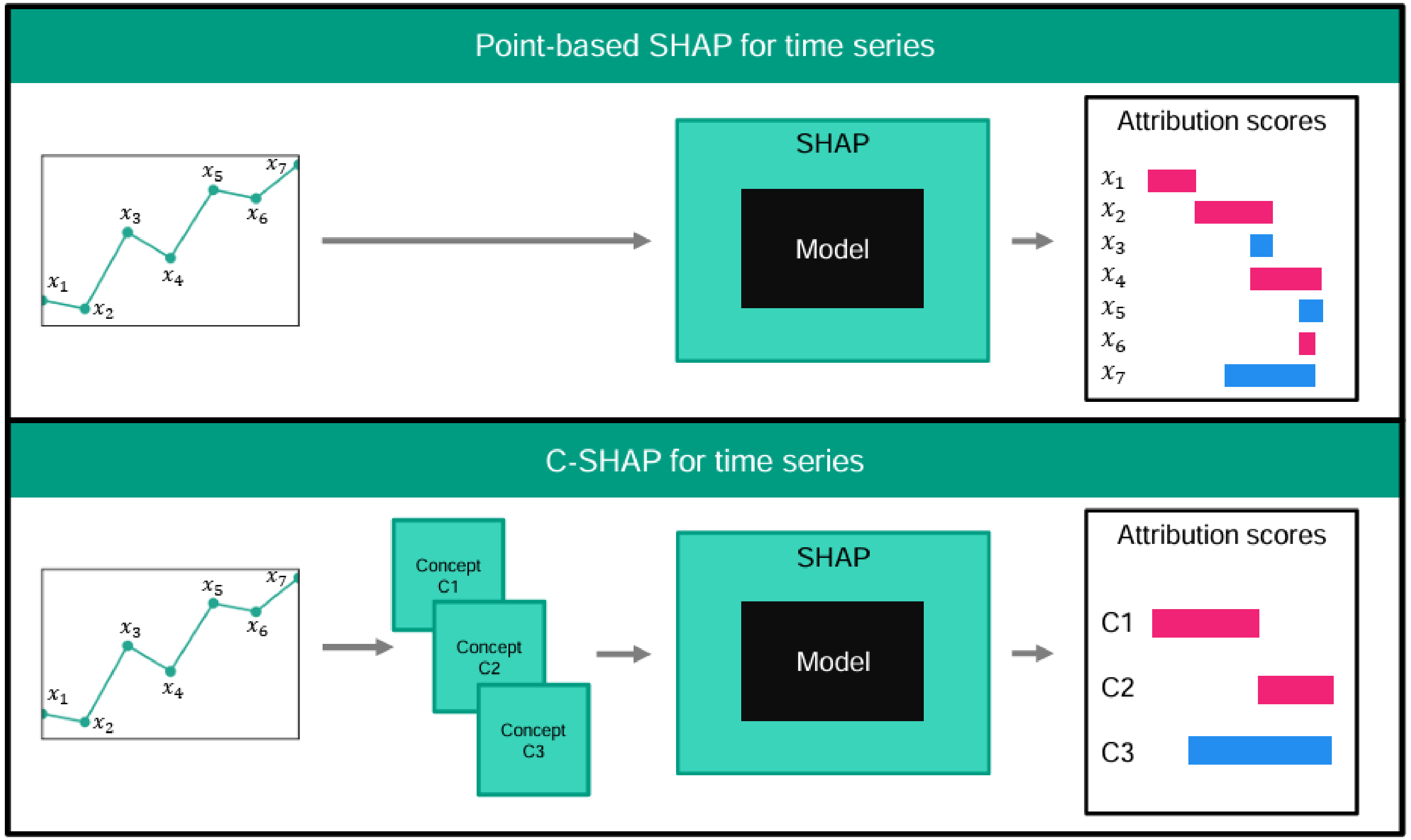


Figure 7. *Heatmaps depicting the importance of all time steps for the important features for a certain patient record from the MIMIC-III dataset. The top 15 variables depicted on the y axis are ranked according to their importance. The darker the color is, the higher the absolute value of the assigned Shapley value is.*



C-SHAP



(https://arxiv.org/abs/2504.11159)

DIGITAL

DIGITAL

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Horizon Europe: Marie Skłodowska-Curie Actions. Neither the European Union nor the granting authority can be held responsible for them.

This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101119635