Explainable Artificial Intelligence on Biosignals for Clinical Decision Support

Lecture Style Tutorial

Miriam Cindy Maurer, Jacqueline Michelle Metsch, Philip Hempel, Theresa Bender, Philip Zaschke, Nicolai Spicher, Anne-Christin Hauschild

Department of Medical Informatics, University Medical Center Göttingen

Content

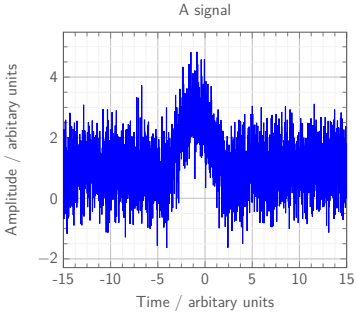
Introduction to Biosignals

Signals - a general introduction

- A signal y is an information-bearing quantity that varies over an independent variable and can be represented as a function x:

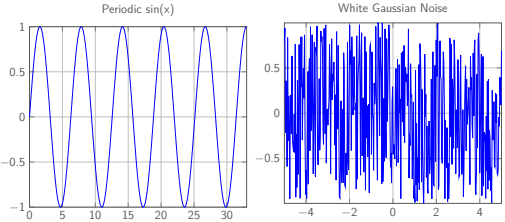
$$y = x(t) \quad x : \mathbb{R} \to \mathbb{C} \quad \text{with } y \in \mathbb{C}, t \in \mathbb{R}$$

- In biosignal research, the independent variable represents time t using the physical unit seconds.
- Due to convention, we might call the depended variables *output*, independent variables *input*, and the number of independent variables *dimensions*.



Deterministic vs. stochastic signals

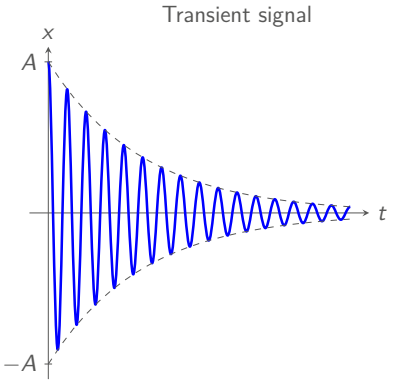
- For *deterministic* signals there is no uncertainty regarding their value at a given point in time. They can be described using equations.
- The exact values of *stochastic* signals for a point in time cannot be predicted and they can only be described statistically.



→ Biosignals oftentimes share properties of both types.

Periodic vs. aperiodic signals

- A signal is called *periodic*, if there exists a fundamental period T_0 such that $x(t) = x(t + mT_0)$, for $m \in$.
- A signal is called *aperiodic*, if no T_0 exists:
  - Aperiodic signals can still be *quasi-periodic*, if they repeat with small errors
  - $\rightarrow$ Many biosignals are quasi-periodic
  - If aperiodic signals do not repeat at all, they are *transient*.
  - $\rightarrow$ Noise in biosignals, e.g. due to motion of the subject.



Periodicity

- Sine and cosine functions are prominent examples of periodic signals.
- The reciprocal of the fundamental period T_0 is the *fundamental frequency* $F_0 = 1/T_0$, measured in unit Hertz [Hz].

Example

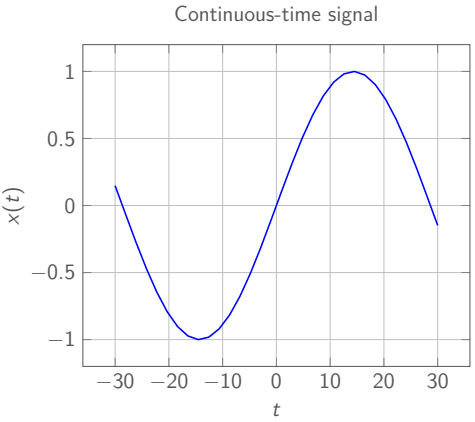
- A heart beats with approx. 1 Hz (60 beats-per-minute) in a calm person.
- The audible frequency range lies within $20 - 20,000$ Hz.
- Red light is an electromagnetic wave with $4 * 10^{14}$ Hz.

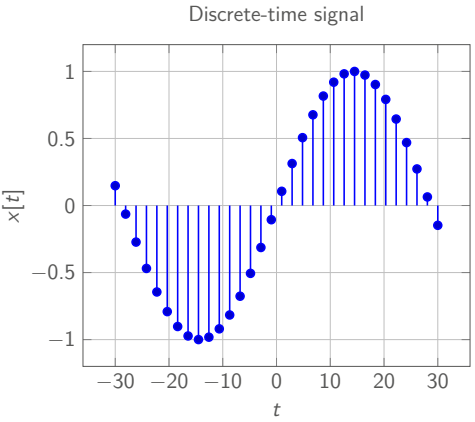
Continuous-time vs. discrete-time signals

- Most discrete signals are acquired by *sampling* a continuous signal, i.e. $x[n] = x(nT_0)$, for $n \in$.
- The number of samples per second is called the *sampling frequency*.

Example

A digital song (e.g. a `.wav` file) is a discrete-time signal $x[t]$. A recorder and microphones were used to store a continuous-time audio wave $x(t)$ in discrete steps (e.g. $T_0 = 1/44.1$ kHz).


Continuous-time signal
1
0.5
0
−0.5
−1
x(t)
−30
−20
−10
0
10
20
30
t


Discrete-time signal
1
0.5
0
−0.5
−1
x[t]
−30
−20
−10
0
10
20
30
t

Biosignals - Attempt of a definition

Definition

Biosignals are signals, that arise from physiological processes in living beings, and that can be continually measured and monitored. Biosignals stem from electrical, mechanical, and chemical changes in the body.

Discuss: Which effects can you measure as a biosigal?

Blood pressure	?
Body temperature	?
Insulin levels	?
Electrical brain activity	?
Blood iron levels	?
Respiration	?

Biosignals - Attempt of a definition

Definition

Biosignals are signals, that arise from physiological processes in living beings, and that can be continuously measured and monitored. Biosignals stem from electrical, mechanical, and chemical changes in the body.

Discuss: Which of the given examples is a biosignal?

Blood pressure	✓
Body temperature	50/50
Insulin levels	X
Electrical brain activity	✓
Blood iron levels	X
Respiration	✓

Nomenclature

Important note

We need to differentiate between:

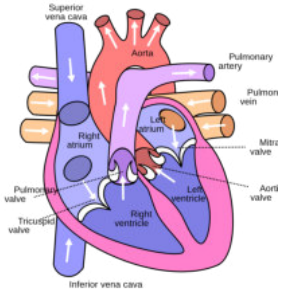
- The physiological entity (e.g. cardiac conduction system)
- The measurement concept (e.g. electrocardiography)
- The measurement device (e.g. patient monitor)
- The measured signal (e.g. electrocardiogram)

Example

We can measure the heart rate frequency, i.e. how often the heart beats, using different modalities: One way is to use the ECG to measure the electrical activity on the chest, another way is to measure the blood flow in the finger via plethysmography.

The human heart

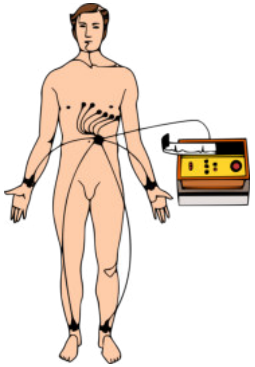
- The heart is a highly specialized organ located in the middle compartment of the chest with a mass of approximately 250 – 350g in an adult and consists mainly of muscle tissue, called myocardium.
- The heart is divided in a left and right half with both consisting of an upper and a lower chamber, called atrium and ventricle, respectively.



By Wapcaplet, Yaddah - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=830246

Electrocardiography

- Electrocardiography (ECG) is a non-invasive measurement of the heart's electrical activity
- 10 electrodes are are attached to the skin and measure electrical potential differences:
  - 6 chest electrodes: $V1 \cdots V6$
  - 4 limb electrodes: Right/left arm/leg
- From these electrodes, 12 "leads" are computed:
  - 6 chest leads: $V1 \cdots V6$ (unipolar)
  - Einthoven:[a] I: RA-LA, II: RA-LL, III: LA-LL
  - Goldberger: aVR: RA/(LA+LB), aVL: LA/(RA+RB), aVF: LB/(LA+RA)

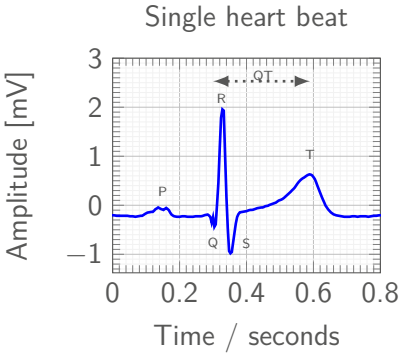


By Madhero88 - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=6098303

[a] W. Einthoven received the Nobel Prize in Physiology or Medicine in 1924 for inventing the first practical ECG device.

Electrocardiography

- The ECG signal of a healthy heartbeat follows a very specific pattern P-QRS-T:
  - P: atrial depolarisation
  - QRS: depolarization of the ventricles
  - T: repolarization of the ventricles
- Voltages are measured in units of [mV] and origin from the depolarization and repolarization of the myocardium.



Electrocardiography irregularities

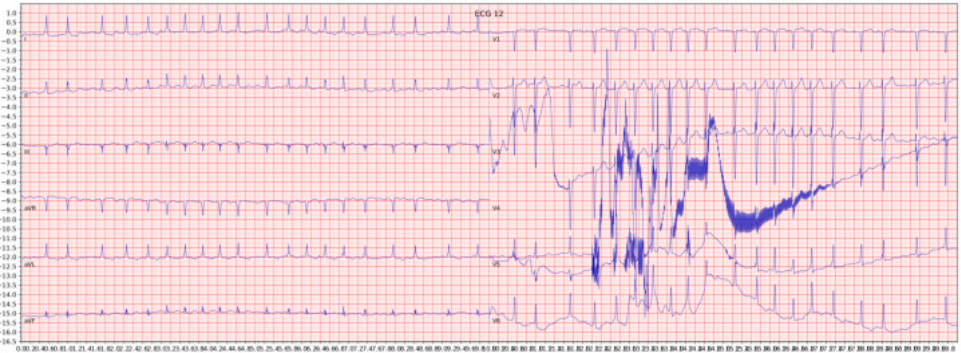
- The ECG enables the detection of a many heart-related conditions, e.g.
  - Myocardial infarction: block in a artery leads to interruption of blood flow
  - Bundle branch blocks: partial or complete interruptions of electrical impulses
  - Arrhythmia: irregular rhythm of the heart beat.
  - Bradycardia/Tachycardia: Abnormally low or high heart rate

Important note

Although the ECG has high diagnostic value, we always have to consider pitfalls:

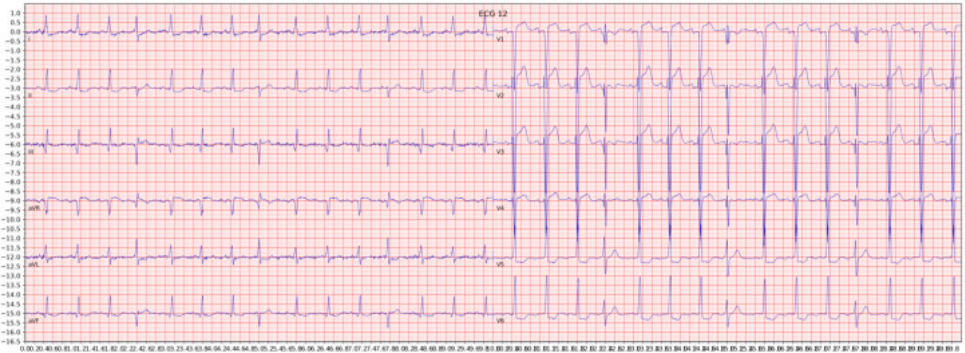
- High degree of intra- and inter-subject variability.
- Measurement noise due to problems with the hardware (e.g. wrong cable setup) or the subject (e.g. motion artifacts)

Electrocardiography examples



Arrythmic ECG with left bundle branch block (LBBB, duration QRS $>$ 120 ms in V5/V6). Heavy noise due to weakly attached V3 / V4 electrodes + muscle noise.

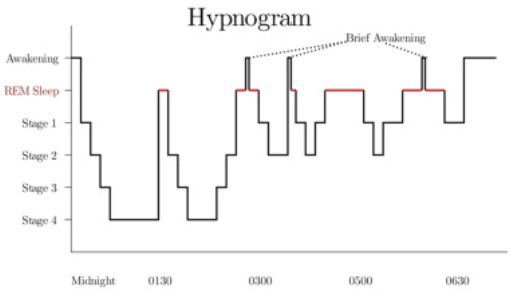
Electrocardiography examples



Myocardial infarct visible as ST elevation (V1-V3) + LBBB

Sleep

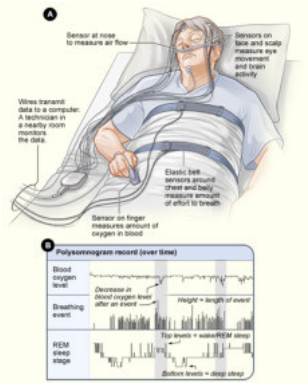
- Sleep follows a typical pattern over night
  - Awake
  - Rapid Eye Movement (REM) Sleep
  - Non-REM sleep
    - N1-N2 Light sleep
    - N3 Deep sleep
- This pattern is distorted in patients suffering from sleep disorders.



By RazerM at English Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17745252

Polysomnography

- For sleep disorder diagnosis, polysomnograms are the gold standard which are acquired in sleep labs
- The standard setup is depicted on the right and includes, next to ECG:
  - Respiratory signals
  - Electroencephalography (EEG): electrical activity of the brain
  - Electrooculography (EOG): electrical activity of the eyes
  - Electromyography (EMG): muscle activity
  - Pulse oximetry

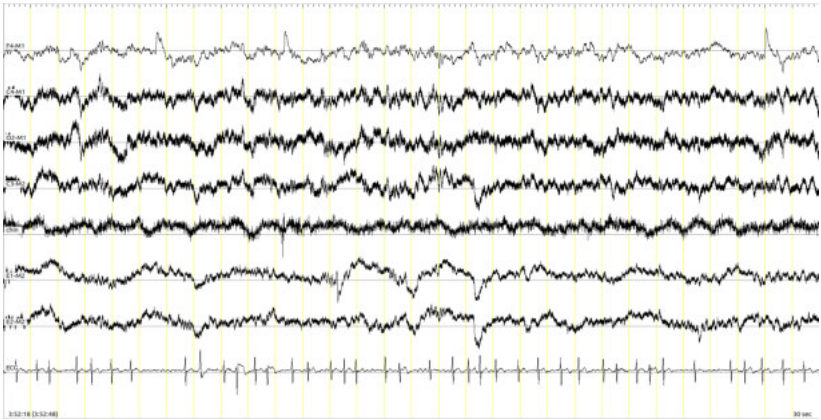


By National Heart Lung and Blood Institute (NIH), Public Domain, https://commons.wikimedia.org/w/index.php?curid=29590111

Polysomnography

- Polyosomnography is acquired from patients with suspected sleep abnormalities, e.g. sleep apnea (repetitive pauses in breathing), narcolepsy (high day drowsiness), or periodic limb movement.
- Data is acquired over a whole night and annotated by humans on the next data in windows of 30 seconds based on the *American Academy of Sleep Medicine (AASM)* reference manual.

Stage	Spindles	Alpha/Theta	Delta	EMG	REM	Slow EM height
W	-	+++	Spikes	+++	+++	+++
N1	-	+	-	++	-	+++
N2	+++	-	+	+	-	+
N3	++	-	+++	-	-	-
R	-	+	-	-	+++	+++



PSG 30s window

Introduction to XAI

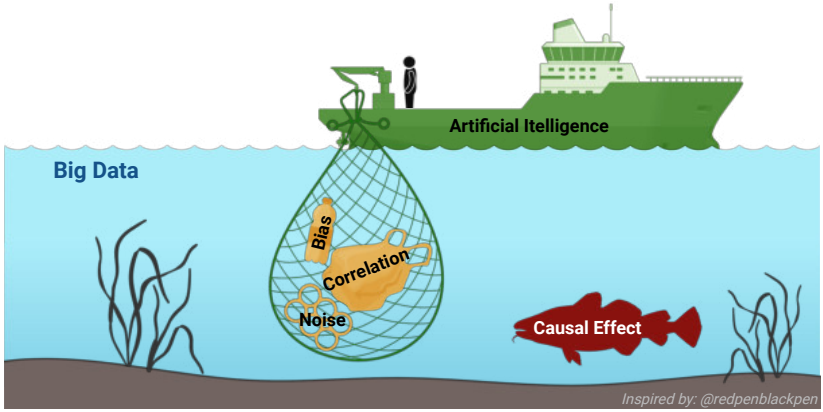
eXplainable AI (XAI)

Discussion

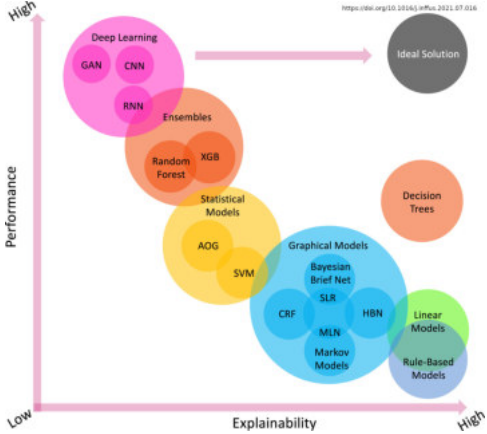
1. Why do we need Explainable AI?
2. How does Explainable AI work?

Why do we need Explainable AI?

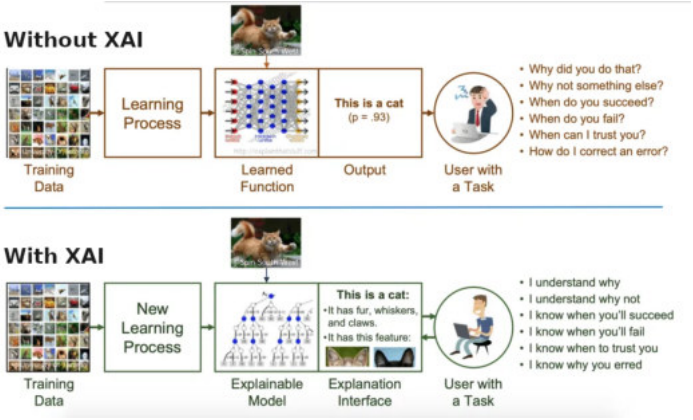
Why do we need Explainable AI?



Why do we need Explainable AI?



Why do we need Explainable AI?



Importance of XAI for Stakeholders

Data Scientists

- Why is XAI important for Data Scientists?

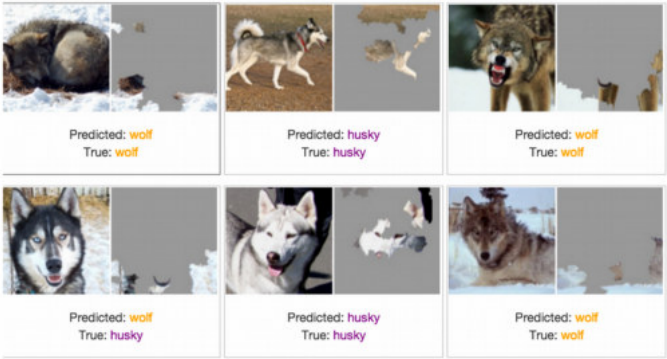
Consumers

- Why is XAI important for Consumers?

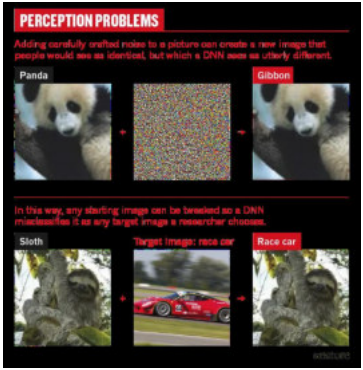
Regulators

- Why is XAI important for Regulators?

Importance of XAI for Data Scientists



Importance of XAI for Data Scientists



Douglas Heaven, Why deep-learning AIs are so easy to fool. Nature NEWS FEATURE. 09 October 2019

Importance of XAI for Stakeholders

Data Scientists

- **Increase Understanding**
- **Improve Performance**
- **Create Better Algorithms**
- **Produce Better Models**

Consumers

- Why is XAI important for Consumers?

Regulators

- Why is XAI important for Regulators?

Importance of XAI for Stakeholders

Data Scientists

- Increase Understanding
- Improve Performance
- Create Better Algorithms
- Produce Better Models

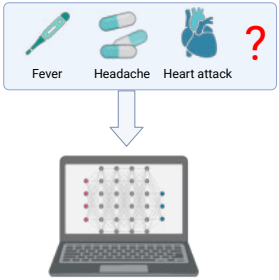
Consumers

- Why is XAI important for Consumers?

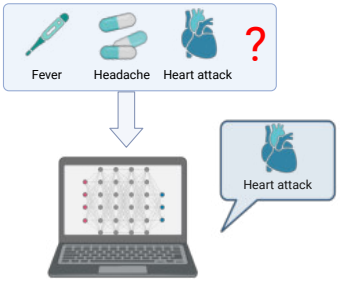
Regulators

- Why is XAI important for Regulators?

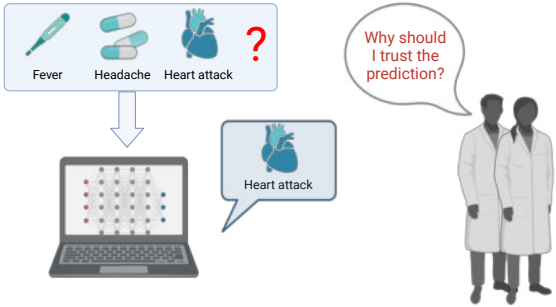
Importance of XAI for Consumers



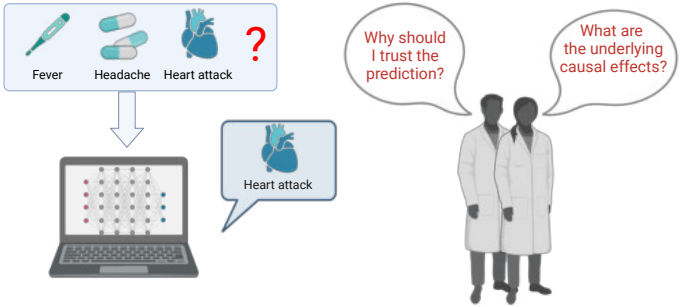
Importance of XAI for Consumers



Importance of XAI for Consumers



Importance of XAI for Consumers



Importance of XAI for Medical Applications

Importance of XAI for Medical Applications

Importance of XAI for Medical Applications

Importance of XAI for Medical Applications

Importance of XAI for Stakeholders

Data Scientists

- Increase Understanding
- Improve Performance
- Create Better Algorithms
- Produce Better Models

Consumers

- **Increase Trust**
- **Bias and Transparency**
- **Understand Impact**
- **Reports and Analysis**

Regulators

- Why is XAI important for Regulators?

Importance of XAI for Stakeholders

Data Scientists

- Increase Understanding
- Improve Performance
- Create Better Algorithms
- Produce Better Models

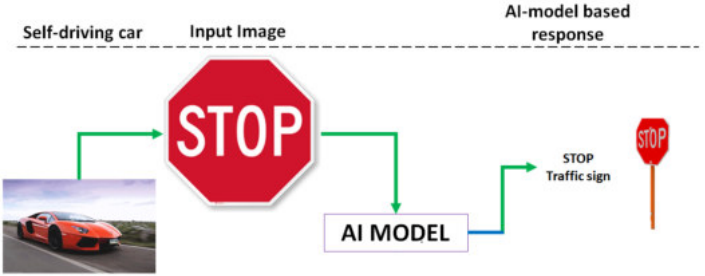
Consumers

- Increase Trust
- Bias and Transparency
- Understand Impact
- Reports and Analysis

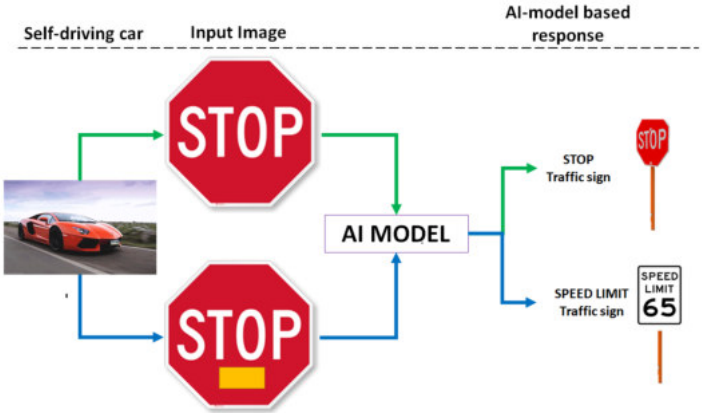
Regulators

- Why is XAI important for Regulators?

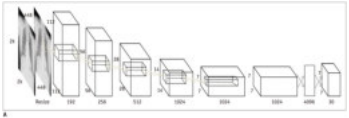
AI in Critical Applications



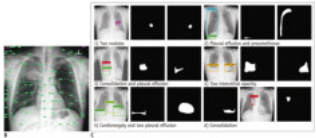
AI in Critical Applications



AI in Critical Applications



Lee, June-Goo, et al. "Deep learning in medical imaging: general overview." Korean journal of radiology 18.4 (2017): 570-584.



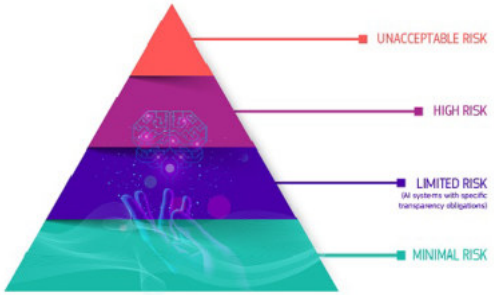
EU Artificial Intelligence Act

https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Legal requirement for:

Algorithm / Model need to support the possibility to retrace why the model made certain decisions.

AI in Critical Applications



High-risk AI systems will be subject to strict obligations before they can be put on the market:

- adequate risk assessment and mitigation systems
- high-quality datasets feeding the system to minimize risks and discriminatory outcomes
- logging of activity to ensure traceability of results
- detailed documentation providing all information necessary on the system and its purpose for authorities to assess its compliance
- clear and adequate information to the user
- appropriate human oversight measures to minimize risk
- high level of robustness, security, and accuracy

https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Importance of XAI for Stakeholders

Data Scientists

- Increase Understanding
- Improve Performance
- Create Better Algorithms
- Produce Better Models

Consumers

- Increase Trust
- Bias and Transparency
- Understand Impact
- Reports and Analysis

Regulators

- **Increase Trust**
- **Bias and Transparency**
- **Compliance**
- **Reports**

Importance of XAI

Summary

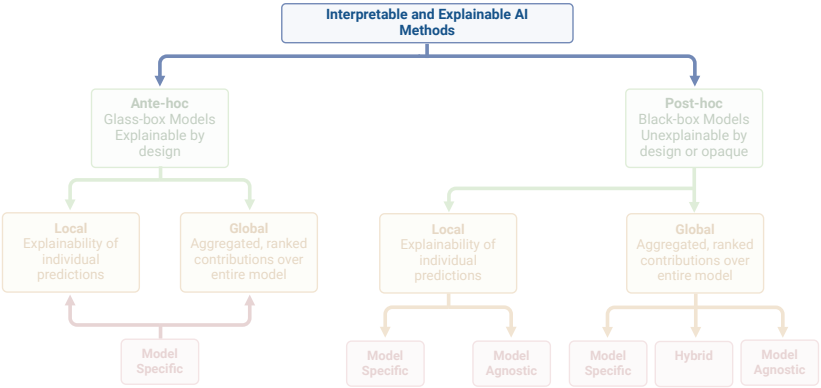
- Explain predictions to support the decision-making process
- Debug the unexpected behavior of a model
- Refine modeling and data collection process
- Verification of model behavior
- Legal aspects
- Retain human control in decision-making
- Establish Trust in the decision process for stakeholders

What makes a good explanation?

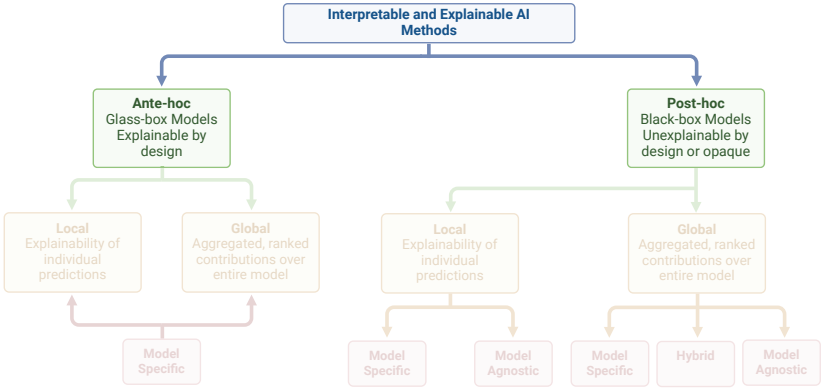
Explanations must be:

- Complete
- Accurate
- Meaningful
- Consistent

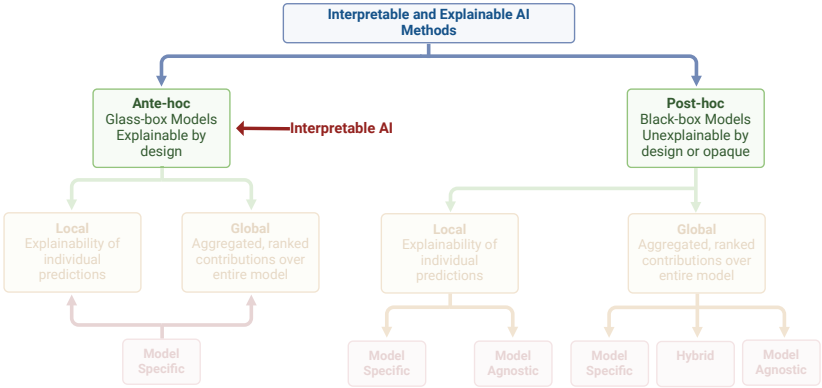
Taxonomy of Interpretable and Explainable AI



Taxonomy of Interpretable and Explainable AI



Taxonomy of Interpretable and Explainable AI



Interpretable Ante-hoc XAI Methods

Ante-hoc

- Interpretable models = Explainable by design
- Importance calculation is directly embedded in the learning algorithm
- "glass-box" models that are ante-hoc interpretable

Interpretable Ante-hoc XAI Methods

Advantages and Disadvantages

Importance calculation is directly embedded in the learning algorithm

Advantages:

- Better runtimes and less complexity
- Dependencies between data points are modeled

Disadvantages:

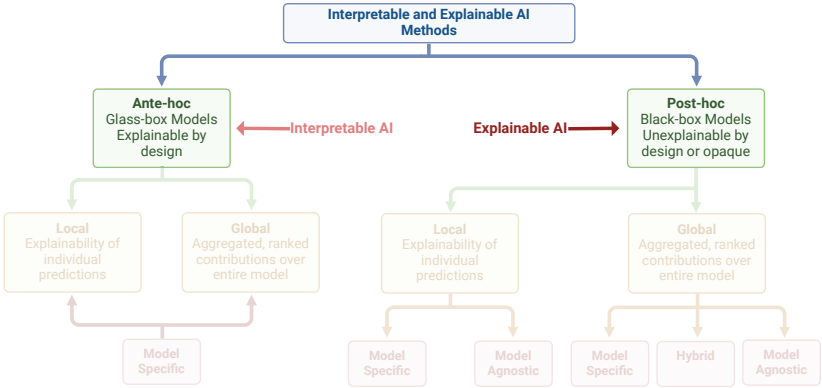
- Strongly depends on the learning algorithm used
- Model-specific biases

Interpretable Ante-hoc XAI Methods

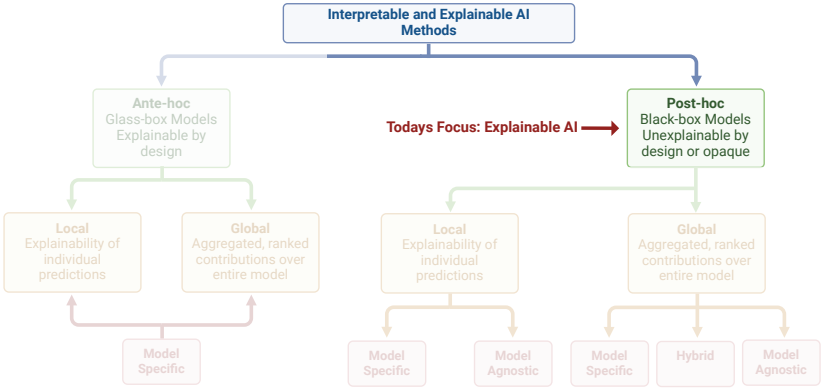
Example

- Linear Methods
  - Logistic Regression: Regression Coefficients
  - (Grouped) Lasso: Penalization towards few features
  - (Regularized) Linear Discriminant Analysis
  - Linear SVM
- Non-Linear Methods
  - Non-Linear Methods
  - Decision Trees
  - Random Forest

Ante-hoc vs Post-hoc Methods



Ante-hoc vs Post-hoc Methods

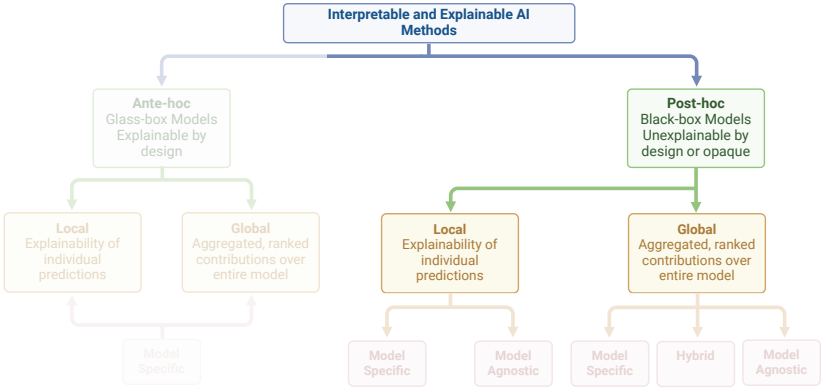


Post-hoc XAI Methods

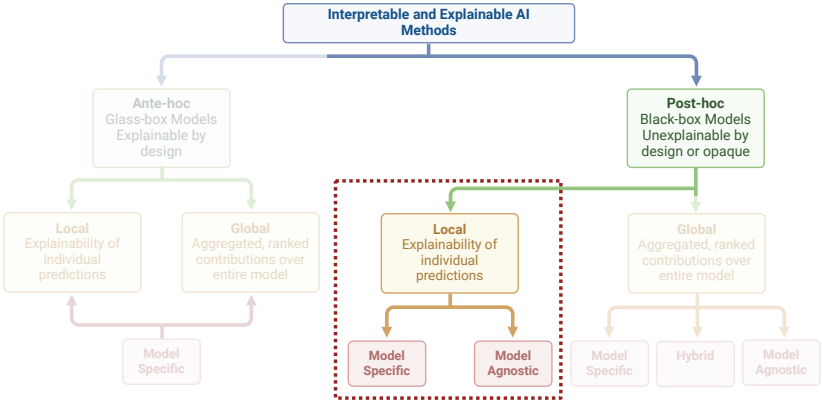
Post-hoc

- Interpreting Black-Box models
- Applying methods that analyze trained models
- Any model can be explained with XAI
- Often applied in imaging based on DL and black box models (Many XAI Methods developed for imaging)

Local vs. Global Explainability



Local vs. Global Explainability

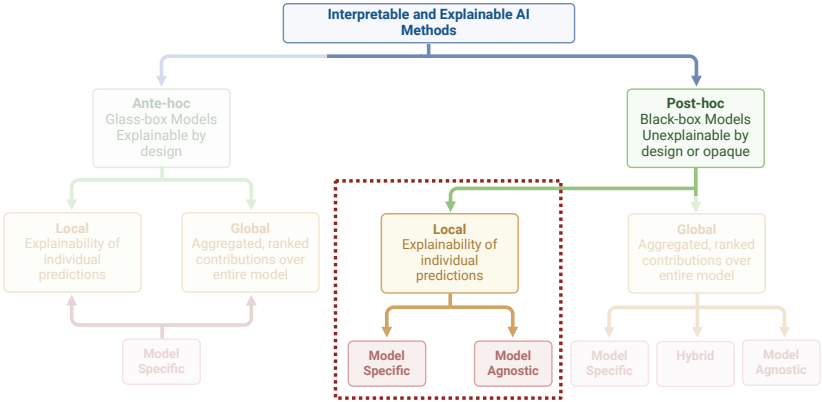


Local Explainable AI

Local

- Interpretation of individual predictions or a small part of the model's prediction space
- "Easy" to understand
- Higher precision but lower recall understanding of model behavior
- Not guaranteed representative since they are calculated for single samples

Model Specific vs. Model Agnostic XAI



Local Model Specific vs. Model Agnostic XA

Model Specific

It only works for specific models due to definition, e.g.

- **Backpropagation-based methods**
- Integrated Gradients
- SmoothGrad

Local Model Specific vs. Model Agnostic XA

Model Specific

It only works for specific models due to definition, e.g.

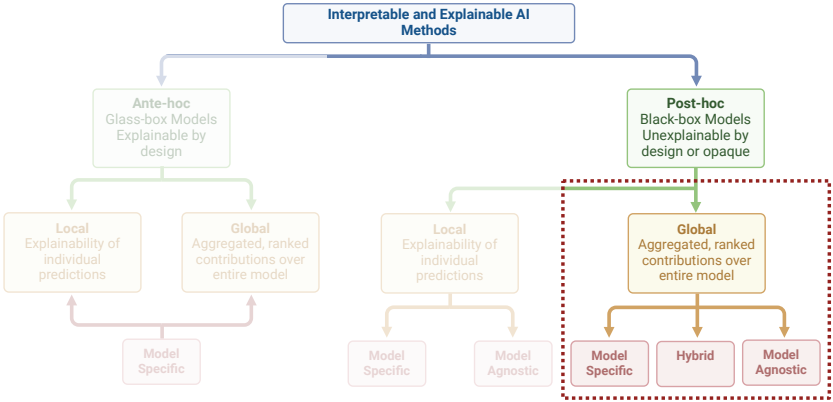
- **Backpropagation-based methods**
- Integrated Gradients
- SmoothGrad

Model Agnostic

Portable across model definitions, e.g.

- **Pertubation-Based Methods**
- **Surrogate Methods, e.g. Local Interpretable Model-agnostic Explanations (LIME)**
- Shapley values (SHAP)

Global Explainable AI



Global Explainable AI

Definition

- Aggregated, ranked contributions of input variables for the entire model's prediction space
- Higher recall view of the entire model prediction space but lower precision due to aggregation, e.g., averages
- Divide the whole input space by set of rules

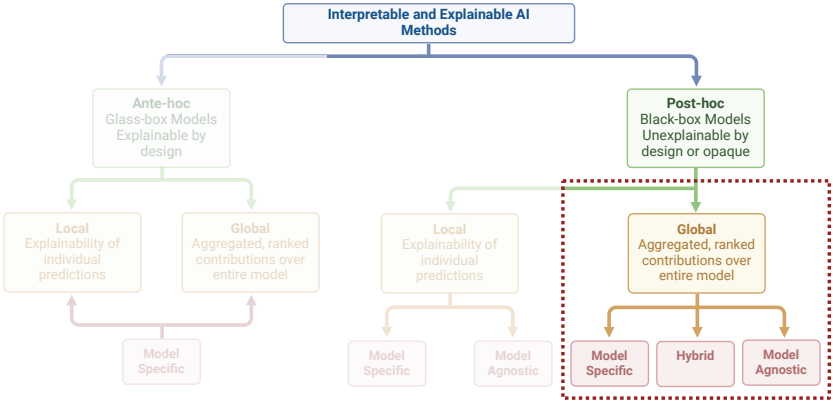
Advantage:

- Are fully transparent and representative of the whole input space

Disadvantage:

- Can quickly be difficult to understand

Global Model Specific vs. Model Agnostic XAI



Global Model Specific vs. Model Agnostic XAI

Model Specific

Only works for specific models due to definition, e.g.

- Tree-based feature importance
- Testing with Concept Activation Vectors (TCAV)

Global Model Specific vs. Model Agnostic XAI

Model Specific

Only works for specific models due to definition, e.g.

- Tree-based feature importance
- Testing with Concept Activation Vectors (TCAV)

Model Agnostic

Portable across model definitions, e.g.

- Permutation feature importance
- Partial Dependence Plots (PDP)

Global Model Specific vs. Model Agnostic XAI

Model Specific

Only works for specific models due to definition, e.g.

- Tree-based feature importance
- Testing with Concept Activation Vectors (TCAV)

Model Agnostic

Portable across model definitions, e.g.

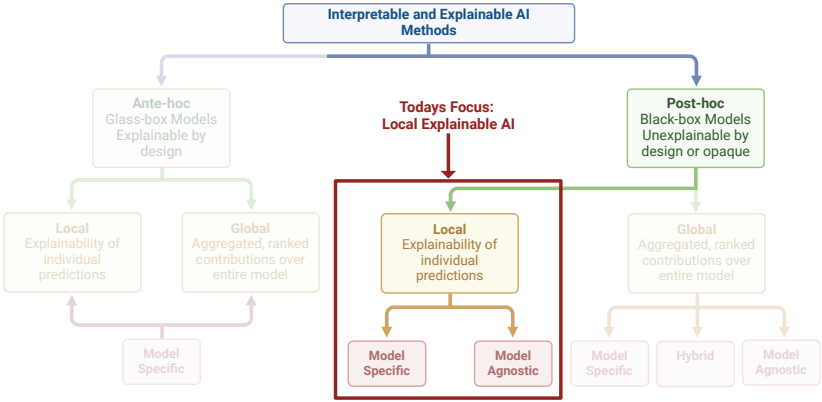
- Permutation feature importance
- Partial Dependence Plots (PDP)

Hybrid

Agreggate Local Explanations e.g.

- SHAP
- Integrated Gradient

Model Specific vs. Model Agnostic XAI



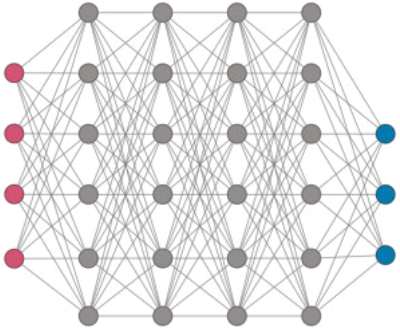
Break (5min)

XAI Workflow

AI Models suited for Biosignals

- Deep Neural Networks
- Convolutional Neural Networks
- Recurrent Neural Networks
- Transformer Networks
- Graph Neural Networks

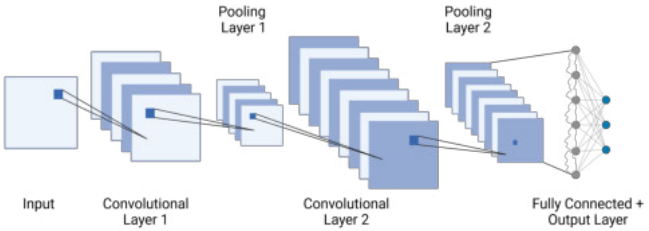
Deep Neural Networks



Feed forward neural network. Image generated with Biorender.

- Able to learn complex relationships in the data
- Only able to take 1D-Input $\rightarrow$ Data needs to be flattened
- Loss of spatial information

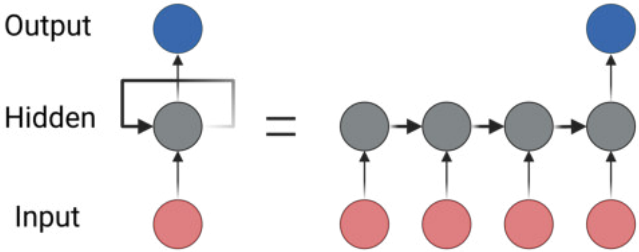
Convolutional Neural Networks



Example of Convolutional Neural Network. Image generated with Biorender.

- Great for pattern recognition (waveforms, shapes, and temporal dynamics)
- Process data while preserving spatial and temporal structures

Recurrent Neural Networks

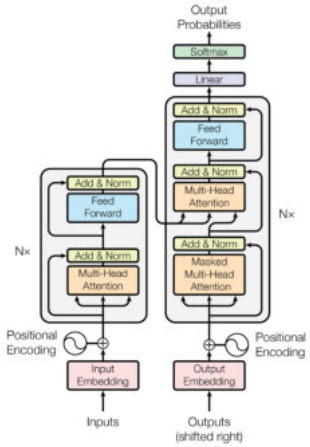


Example of Recurrent Neural Network. Image generated with Biorender.

- Capture dependencies in time-series data effectively
- Especially Long Short Term Memory Networks (LSTMs) can carry information for long periods of time

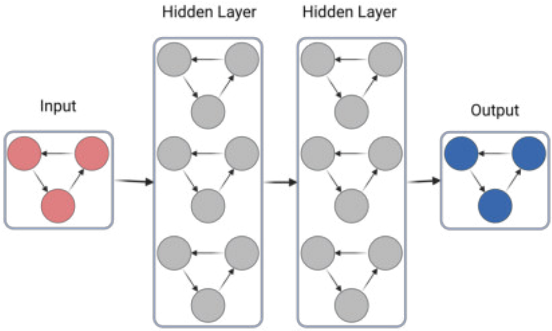
Transformer Networks

- Ability to focus on specific parts of the signal that are more informative for the task
- Can handle entire sequences at once
- Can process and relate distant points in the sequence directly without step-wise propagation



The Transformer architecture. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Graph Neural Networks



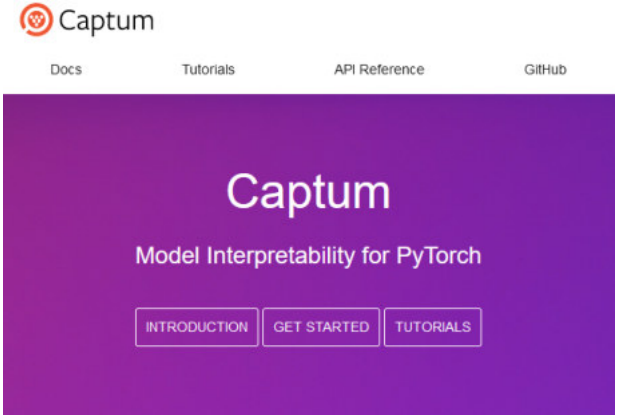
Example of Graph Neural Network. Image generated with Biorender.

- Can model the complex relationships and structures within data represented as graphs (e.g. brain connectivity networks from EEG)
- Can incorporate spatial information about the nodes, which is particularly useful for biosignals recorded from spatially distributed sensors

Generating XAI Attributions

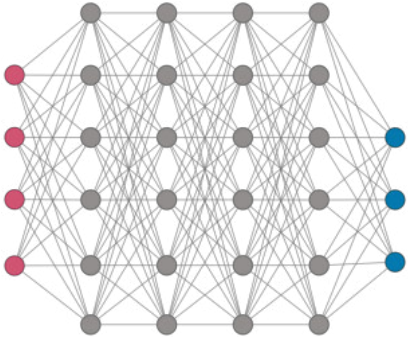
- Captum AI
- Backpropagation Based Methods
- Pertubation Based Methods
- Surrogate Methods

Captum AI



Screenshot of thr Captum AI website: https://captum.ai/.

Backpropagation Based Methods



- Track the gradients of the output (e.g., classification scores) with respect to the input features through the whole network

Backpropagation Based Methods

- Based on the intuition that if a small change in an input feature significantly affects the output, that feature must be important for the decision-making process
- Can sometimes produce misleading or difficult-to-interpret explanations, especially in highly complex networks
- Examples: Input$\times$Gradient, Integrated Gradients, LRP, GradCAM, Deconvolution, Guided Backpropagation, and DeepLIFT

Pertubation Based Methods

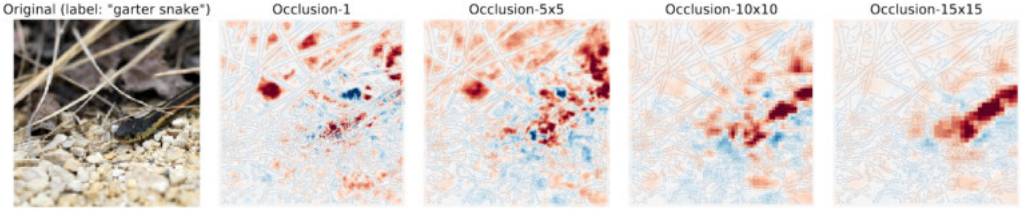


Image from Ancona, Marco, et al. "Towards better understanding of gradient-based attribution methods for deep neural networks." arXiv preprint arXiv:1711.06104 (2017)

- Interpret the decisions by systematically altering (perturbing) the input data and observing the impact on the model's output

Pertubation Based Methods

- They are not dependent on underlying model parameters
- Can be computationally intensive, especially for large datasets or complex models
- The impact of perturbations can sometimes be difficult to interpret, especially for interactions between features
- Examples: Occlusion, Feature Permutation, and Shapley Values

Surrogate Methods

- Use a simpler model to emulate the predictions of a complex model across a specific input space or dataset
- By training the surrogate model to approximate the outputs of the original model, one can analyze the surrogate model to gain insights into how the original model makes decisions
- The interpretability of the surrogate model allows for the extraction of human-understandable explanations, such as feature importance, decision rules, or visualizations.
- Examples: LIME

Break (5min)

Visualizing Relevance Attributions

Visualizing Relevance Attributions

- Challenges in visualizing ECG relevance attributions
- Solutions and adaptations for clinical relevance

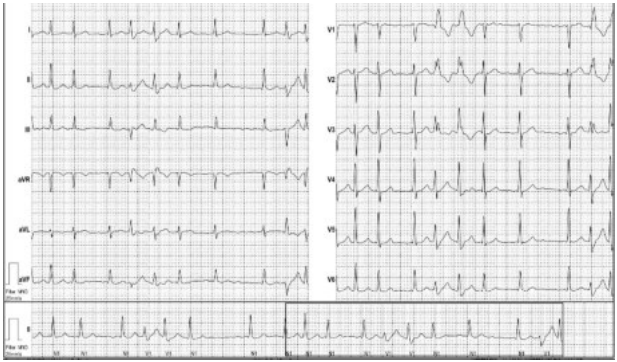


Figure: Real World Example; Source X: Andreas Roeschl 30.06.24

Challenges and Solutions in Visualization

- Challenges:
  - Ensuring clinical relevance of visualizations
  - Normalizing relevance attributions effectively
  - Bridging the gap between AI models and cardiologists
- Solutions:
  - Normalization using quartiles for robust relevance scaling
  - Customizable color schemes to indicate disease relevance
  - Incorporating standard ECG grid to enhance interpretability

Importance of ECG Grid for Cardiologists

- ECG grid provides a standardized way to interpret ECG signals
- Essential for identifying patterns, anomalies, and diagnosing conditions
- Our function adapts this by:
  - Setting up x and y ticks to match clinical standards
  - Customizing grid lines for clarity
  - Overlaying AI relevance attributions on the standard ECG grid
- Enhances trust and usability of AI-assisted diagnostics

Figure: Example of an XAI visualization technique: A neural network trained to detect a right bundle branch block (RBBB) in a standard 12-lead ECG has correctly classified a healthy patient as "no RBBB". The blue-green color highlights negative relevances that contradict the presence of RBBB on the leads, medically referred to as "narrow QRS complexes". This indicates that the decision-making process of the neural network is focused on the same regions of the ECG described in the medical guideline.

Evaluating Relevance Attributions

- Introduction to evaluation metrics
- Importance of qualitative and quantitative evaluation
- Linking relevance to medical guidelines
- Example: ECG and clinical validation
- Tools and frameworks for evaluation

Evaluation Metrics

- Qualitative vs. quantitative evaluation
- Common metrics: accuracy, precision, recall
- Specific metrics for relevance attributions
- Challenges in evaluation

Qualitative Evaluation

- Individual subject analysis
- Visual inspection of relevance maps
- Case studies and expert validation

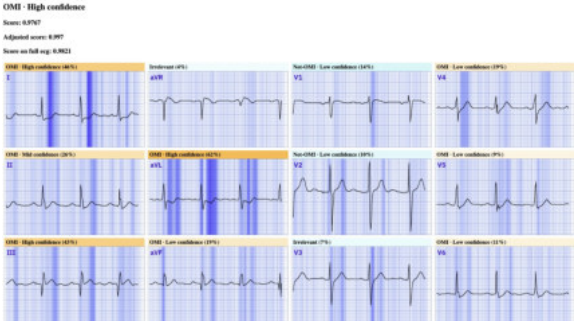


Figure: Case study: 40sM presents to ER from work w atypical CP (dull/pressure, 8/10, but worse w palpitation) radiating to the jaw. ER sends this EKG. Source X: Robert Herman

Quantitative Evaluation

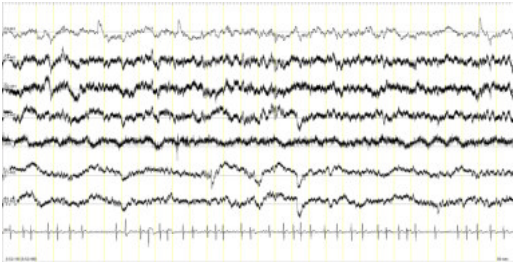
- Analysis of relevance data
- Aggregating relevance across subjects
- Correlation with clinical outcomes
- Example: cohort analysis in ECG studies

Use Cases

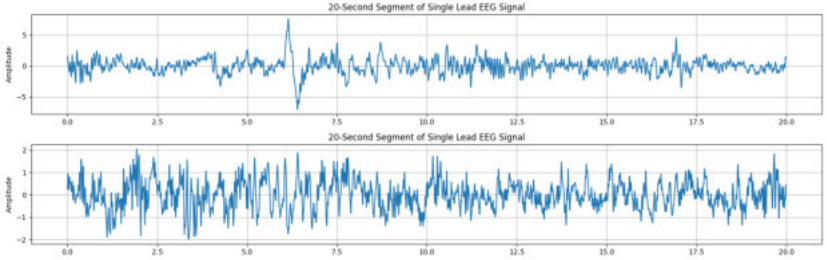
EEG - Sleep Stage Classification

Using 1-channel EEG

- Polysomnography contains multiple signals, such as ECG, EOG, EEG
- We will focus now on the brain waves only for sleep stage classification



EEG examples



How to differentiate sleep stages?

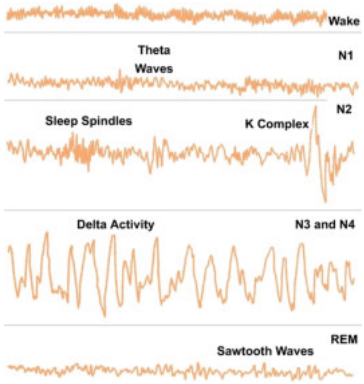
- Specific frequencies particularly strong
- Morphological features present (spindles, spikes)

Stage	Spindles	Alpha/Theta	Delta	EMG	REM	Slow EM height
W	-	+++	Spikes	+++	+++	+++
N1	-	+	-	++	-	+++
N2	+++	-	+	+	-	+
N3	++	-	+++	-	-	-
R	-	+	-	-	+++	+++

Sleep Stage Characteristics

Figure: EEG characteristics of the different sleep stages[a]

[a] Dutt et al. 2023.



Data

- Sleep Heart Health Study (SHHS)[1] - first visit
- 5,793 polysomnography recordings
- C3/A2 and C4/A1 EEGs, sampled at 125 Hz
- Sleep stages (W, R, N1-3) were annotated manually by a central polysomnography reading center

[1] Quan et al. 1997.

Methods

- Adaptation and training of DRCNN[2]
- Post-hoc XAI method Integrated Gradients (IG) from CaptumAI[3]
- Visualization of XAI methods on all samples

[2] Howe-Patterson, Pourbabaee, and Benard 2018.
[3] Kokhlikyan et al. 2020.

Jupyter Notebook

Results

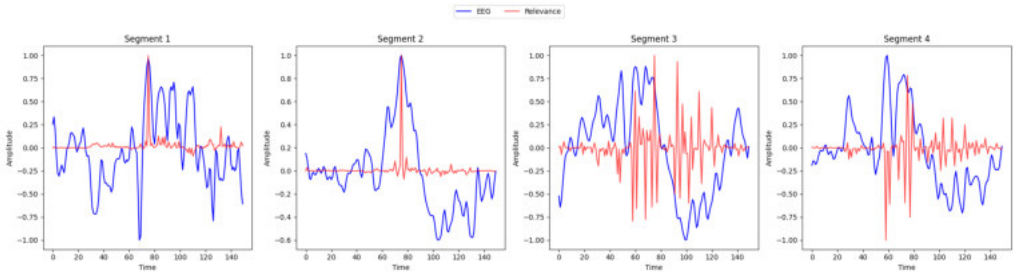


Figure: Example of 1-channel EEG segments during N2 sleep (blue). The relevances using the Integrated Gradients method are plotted in red for the importance of the sample to the neural network.

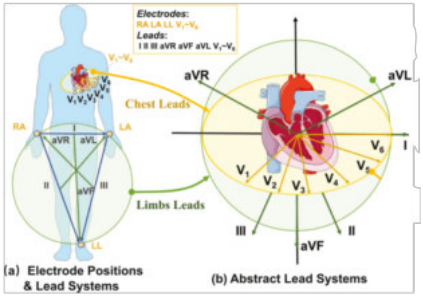
Use Cases

ECG - Detection of RBBB

Hunting Bunnies

Why do we want to use AI for ECG data?

- Deep neural networks show promising results in detecting cardiovascular diseases using 12-lead ECGs
- Providing explanations is crucial
- Relevance attributions of post-hoc XAI methods can be visualized using heatmaps



How to find Right Bundle Branch Block(RBBB)?

- Heartbeat is out of sync
- Particularly visible in leads V1-V3
- Two distinct R-peaks, called "bunny ears"

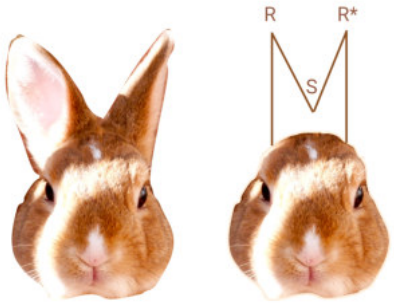
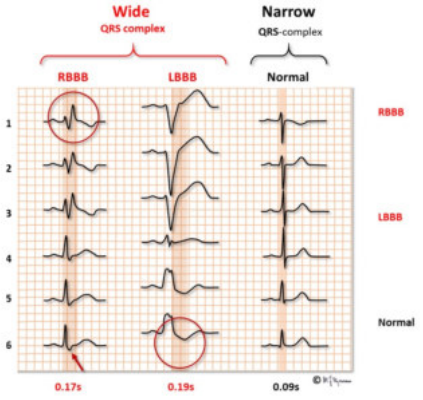
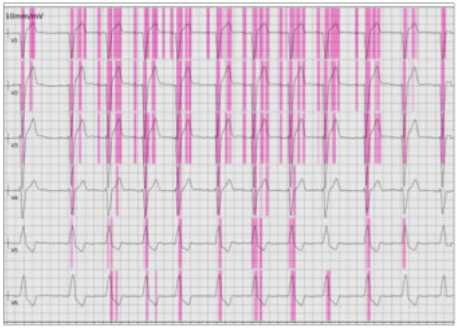


Figure: Characteristic ECG abnormality for the detection of RBBB, namely RSR* pattern

Why did we choose RBBB for this analysis?





Data

- CPSC subset PhysioNet-Computing in Cardiology Challenge 2020[4]
- 6,877 ECGs
- 5,020 cases without RBBB, 1,857 cases with RBBB
- Denoised by Turbé et al.[5] using Empirical mode decomposition for low and high frequency artefacts

[4] Liu et al. 2018.
[5] Turbé et al. 2023a.

Methods

- Pretrained CNN by Turbé et al.[6]
- 15 post-hoc XAI methods from CaptumAI[7]
- Visualization of XAI methods on all samples

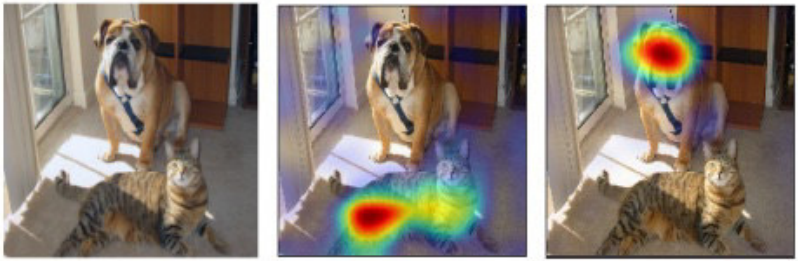
[6] Turbé et al. 2023b.
[7] Kokhlikyan et al. 2020.

What XAI Methods were used?

- Backpropagation based Methods:
  Input $\times$ Gradient; Deconvolution; Guided Backpropagation; Integrated Gradients; LRP; $LRP-\epsilon$; $LRP-\alpha-1-\beta-0$; $LRP-\gamma$; DeepLIFT
- Pertubation based Methods:
  Occlusion; Shapley value sampling; GradientSHAP; KernelSHAP; DeepSHAP
- Surrogate Methods:
  LIME

Why do we want to compare XAI methods?

- No standard taxonomy for all explainable algorithms
- Performance of XAI methods depends on underlying dataset and structural nuances of the model



Results

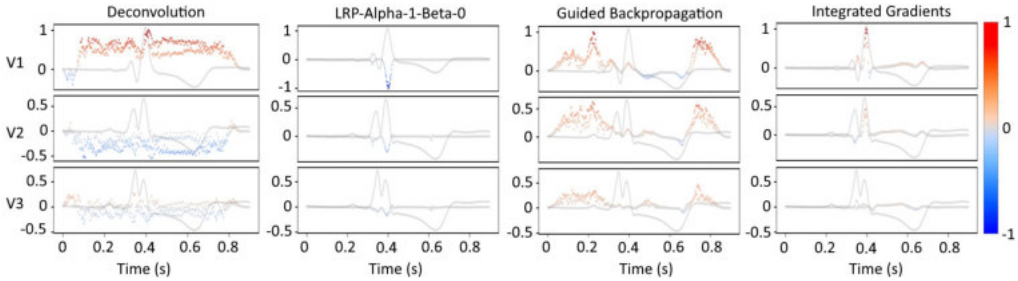


Figure: Average beats of leads V1-V3 (ECG ID: A6800). Relevance attributions by Deconvolution,LRP-α-1-β-0, Guided Backpropagationand Integrated Gradients method (left to right) are plotted directly onto the signal and colored to see, where in the signal the relevant information lies, according to each method.

Jupyter Notebook

Conclusion

- Different XAI methods yield different results
- Difficult to decide whether the model have learned the characteristic ”bunny ears”
- Use multiple XAI methods to increase trust in the predictions

Workshop Wrap Up

- A brief introduction to biosignals and XAI was given
- The importance of XAI in biosignal data was highlighted
- A workflow for the implementation and visualisation of XAI in the clinical field was presented
- Integration of AI and XAI into clinical practice was developed in two use cases with corresponding Jupyter notebooks

Figure: DALL-E

Dutt, Micheal et al. (July 2023). "SleepXAI: An explainable deep learning approach for multi-class sleep stage identification". en. In: *Applied Intelligence* 53.13, pp. 16830–16843. ISSN: 1573-7497. DOI: 10.1007/s10489-022-04357-8. URL: https://doi.org/10.1007/s10489-022-04357-8 (visited on 07/09/2024).

Howe-Patterson, Matthew, Bahareh Pourbabaee, and Frederic Benard (Sept. 2018). "Automated Detection of Sleep Arousals From Polysomnography Data Using a Dense Convolutional Neural Network". In: *2018 Computing in Cardiology Conference (CinC)*. Vol. 45. ISSN: 2325-887X, pp. 1–4. DOI: 10.22489/CinC.2018.232. URL: https://doi.org/10.22489/CinC.2018.232 (visited on 07/09/2024).

Kokhlikyan, Narine et al. (2020). *Captum: A unified and generic model interpretability library for PyTorch*. arXiv: 2009.07896 [cs.LG].

Liu, Feifei et al. (Sept. 2018). "An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection". In: *Journal of Medical Imaging and Health Informatics* 8, pp. 1368–1373. DOI: 10.1166/jmihi.2018.2442.

Quan, S. F. et al. (1997). "The Sleep Heart Health Study: design, rationale, and methods". In: *Sleep* 20.12, pp. 1077–1085. ISSN: 0161-8105.

Turbé, Hugues et al. (Jan. 2023a). *Dataset: Evaluation of post-hoc interpretability methods in time-series classification*. DOI: 10.5281/zenodo.7534770. URL: https://doi.org/10.5281/zenodo.7534770.

— (Mar. 2023b). "Evaluation of post-hoc interpretability methods in time-series classification". In: *Nat. Mach. Intell.* 5.3, pp. 250–260. DOI: 10.1038/s42256-023-00620-w. URL: https://doi.org/10.1038/s42256-023-00620-w.

Input $\times$ Gradient

For an input x Input$\times$Gradient is calculated by

$$r_i^{(\ell)} := x_i \cdot \frac{\partial F}{\partial x_i^{(\ell)}}$$

with $\frac{\partial F(x)}{\partial x_i}$ the gradient of $F(x)$ along the i-th dimension.

Integrated Gradients

We consider the straightline path (in ^{n}) from baseline x' to input x and compute the gradients at all points along the path. Integrated gradients are obtained by cumulating these gradients. Specifically, integrated gradients are defined as the path integral of the gradients along the straightline path from baseline x' to input x.

$$\mathrm{IG}_i(x) := (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

with $\frac{\partial F(x)}{\partial x_i}$ the gradient of $F(x)$ along the i-th dimension.

LRP

Let ℓ and $\ell + 1$ be two consecutive layers in a neural network and $x_i^{(\ell)}$ and $x_j^{(\ell+1)}$ two neurons from layers ℓ and $\ell + 1$ respectively with weighted connection $\omega_{ij}^{(\ell,\ell+1)}$ and bias term $b_j^{(\ell+1)}$.

The relevance attribution $r_i^{(\ell)}$ of neuron $x_i^{(\ell)}$ from layer (ℓ) is calculated as follows:

$$r_i^{(\ell)} = \sum_{j \in \ell+1} r_j^{(\ell+1)} \frac{x_i^{(\ell)} \omega_{ij}^{(\ell,\ell+1)}}{\sum_{\tilde{i} \in \ell} x_{\tilde{i}}^{(\ell)} \omega_{\tilde{i}j}^{(\ell,\ell+1)} + b_j^{(\ell+1)}}.$$

Deconvolution

For an input x the relevance attribution $r_i^{(\ell)}$ is calculated as follows (using the chain rule)

$$r_i^{(\ell)} := x_i^{(\ell)} \cdot \frac{\partial F}{\partial x_i^{(\ell)}} = x_i^{(\ell)} \sum_{j \in (\ell+1)} \frac{\partial F}{\partial x_j^{(\ell+1)}} \frac{\partial x_j^{(\ell+1)}}{\partial x_i^{(\ell)}},$$

if there is no ReLU function in layer (ℓ). For a layer (ℓ) with ReLU activation in the forward pass, the relevance attribution is calculated as follows

$$r_i^{(\ell)} := x_i^{(\ell)} \sum_{j \in (\ell+1)} \frac{\partial F}{\partial x_j^{(\ell+1)}} \text{ReLU}\left(r_j^{(\ell+1)}\right),$$

so only positive relevances are backpropagated.

Guided Backpropagation

For an input x the relevance attribution $r_i^{(\ell)}$ is calculated as follows (using the chain rule [])

$$r_i^{(\ell)} := x_i^{(\ell)} \cdot \frac{\partial F}{\partial x_i^{(\ell)}} = x_i^{(\ell)} \sum_{j \in (\ell+1)} \frac{\partial F}{\partial x_j^{(\ell+1)}} \frac{\partial x_j^{(\ell+1)}}{\partial x_i^{(\ell)}},$$

if there is no ReLU function in layer (ℓ). For a layer (ℓ) with ReLU activation in the forward pass, the relevance attribution is calculated as follows

$$r_i^{(\ell)} := x_i^{(\ell)} \sum_{j \in (\ell+1)} \frac{\partial F}{\partial x_j^{(\ell+1)}} \frac{\partial x_j^{(\ell+1)}}{\partial x_i^{(\ell)}} \text{ReLU}\left(r_j^{(\ell+1)}\right).$$

DeepLIFT

Let ℓ and $\ell + 1$ be two consecutive layers in a neural network and $x_i^{(\ell)}$ and $x_j^{(\ell+1)}$ two neurons from layers ℓ and $\ell + 1$ respectively with weighted connection $\omega_{ij}^{(\ell,\ell+1)}$. Let $\bar{x}_i^{(\ell)}$ be the activation of neuron $x_i^{(\ell)}$ for a baseline input.

The relevance attribution $r_i^{(\ell)}$ of neuron $x_i^{(\ell)}$ from layer (ℓ) is calculated as follows:

$$r_i^{(\ell)} = \sum_{j \in \ell+1} r_j^{(\ell+1)} \frac{x_i^{(\ell)} \omega_{ij}^{(\ell,\ell+1)} - \bar{x}_i^{(\ell)} \omega_{ij}^{(\ell,\ell+1)}}{\sum_{\tilde{i} \in \ell} x_{\tilde{i}}^{(\ell)} \omega_{\tilde{i}j}^{(\ell,\ell+1)} - \bar{x}_{\tilde{i}}^{(\ell)} \omega_{\tilde{i}j}^{(\ell,\ell+1)}}.$$

Occlusion

For tabular data this method masks one feature in the input, runs a forward pass trough the network and then computes the difference between the outputs as the relevance attribution. For image data this can be done for each pixel individually, or by selecting a window size that slides over the image and sets all pixels underneath the window to zero and then calculating the difference in outputs.

Shapley Values

Let F be the set of all features. A prediction is a superadditive map $\nu : 2^F \to$ satisfying $\nu(\emptyset) = 0$. Given a prediction ν, the Shapley value for a feature i is

$$\Phi_i(\nu) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left(\nu(S \cup \{i\}) - \nu(S)\right).$$

LIME

Let $f : ^d \rightarrow$ denote the model being explained and $g \in G$ denote an interpretable model, e.g. linear model, decision tree, with complexity $\Omega(g)$, where G is a class of interpretable models. Further, let $\pi_x(z)$ be a proximity measure between an instance z to x, defining a locality around x. $\mathcal{L}(f, g, \pi_x)$ denotes a fidelity measure of how unfaithful g is in approximating f in the locality defined by π_x, e.g. mean squared error.
The explanation produced by LIME is obtained by the following minimisation:

$$\text{LIME}(x) = \text{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$