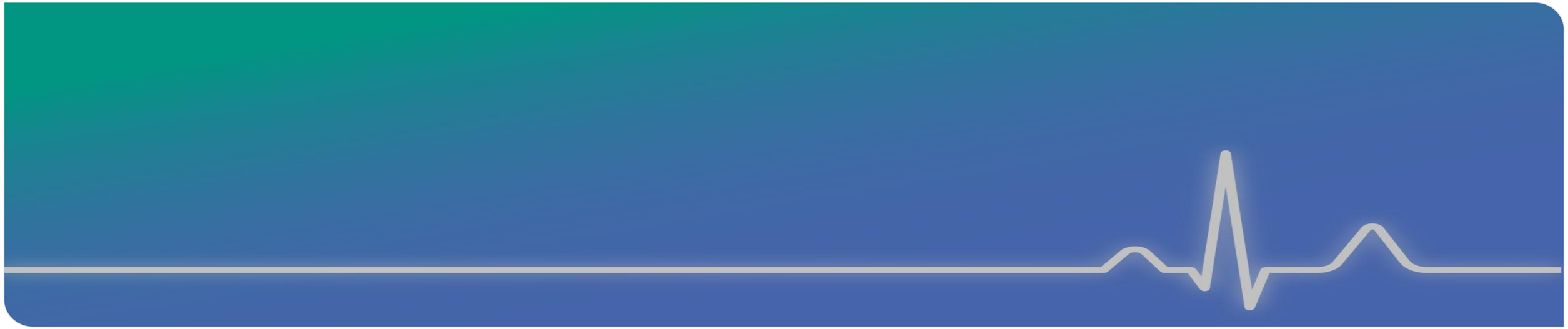
Karlsruhe Institute of Technology

Artificial Intelligence and Machine Learning in Medicine

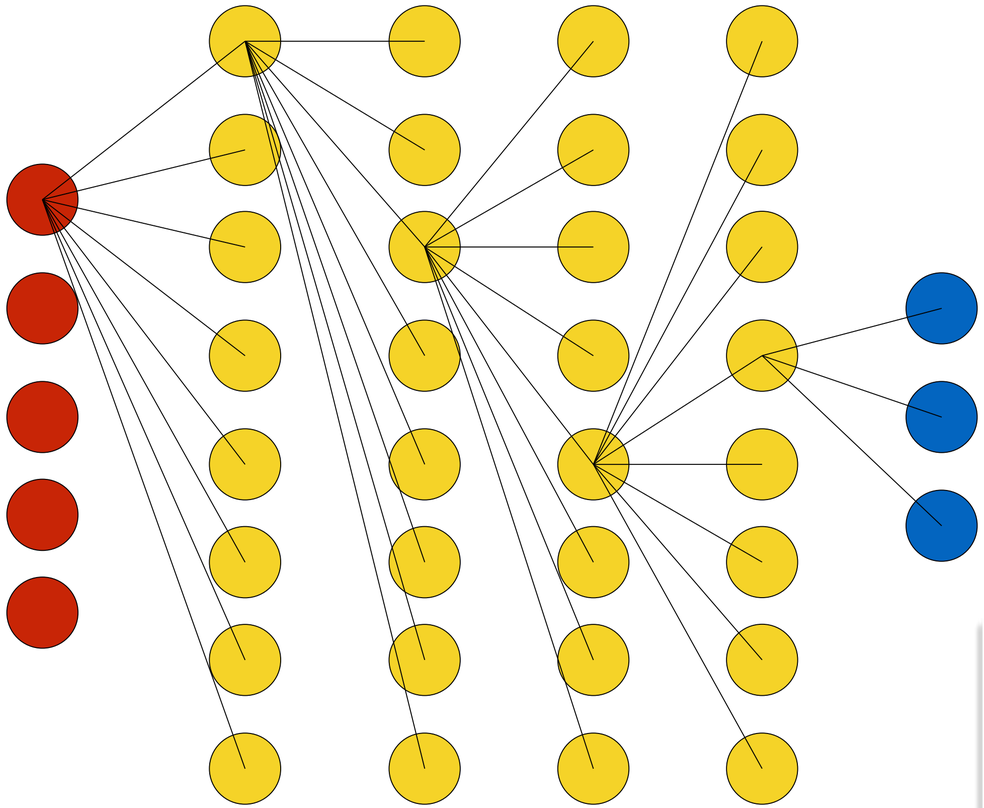
Olaf Dössel, Institute of Biomedical Engineering, Karlsruhe Institute of Technology



Artificial Intelligence and Machine Learning in Medicine

- medical images (X-ray, CT, MRI,…)
- medical measurements (ECG, blood pressure,…)
- in-vitro diagnostics (blood, urine…)
- genetic data (genetic snippets, …genetic code)

Deep Neural Network



input layer

hidden layer

output layer

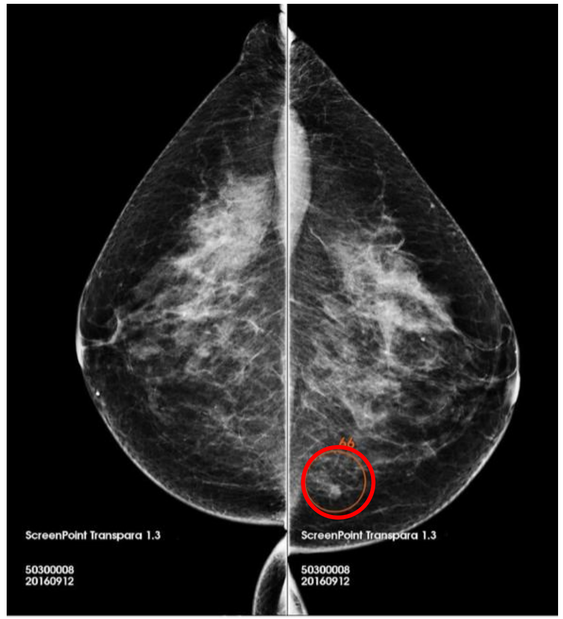
- prognosis / predisposition
- hidden diagnostic properties
- proposal of diagnosis
- proposal of therapy

- Polynomial Regression
- Support Vector Machine
- Decision Tree / Random Forest
- k Nearest Neighbor

Artificial Intelligence and Machine Learning in Medicine

- Some Examples and 3 Projects at IBT
- Some Basics of ML
- How can we Measure the Quality of an AI / ML Algorithm
- Ethical Aspects
- Regulatory Aspects
- More Data for Research - Data Protection and Reference Data
- Some Statements

Mammography Screening



Is the error rate of “single reading plus CAD/ML” lower than with “double reading”?

CAD=Computer Assisted Diagnosis

In 2017 we had 2.8 million examinations within the Mammography screening program in Germany. Double reading by 2 experts is mandatory.

J Med Screen 2021 Sep;28(3):365-368. doi: 10.1177/0969141320984198. Retrospective comparison between single reading plus an artificial intelligence algorithm and two-view digital tomosynthesis with double reading in breast screening
Axel Graewingholt, Stephen Duffy

Detection of Atrial Fibrillation

The NEW ENGLAND JOURNAL of MEDICINE

Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation

Marco Perez et al.
Stanford Center for Clinical Research
14. November 2019

- optical heart beat sensor
- 419 297 participants
- 2161 (0.52%) received notification of irregular pulse
- 450 of them received an ECG patch
- 35% of them had Atrial Fibrillation

funded by Apple

This watch keeps an eye on your heart.

This can only work with Machine Learning!

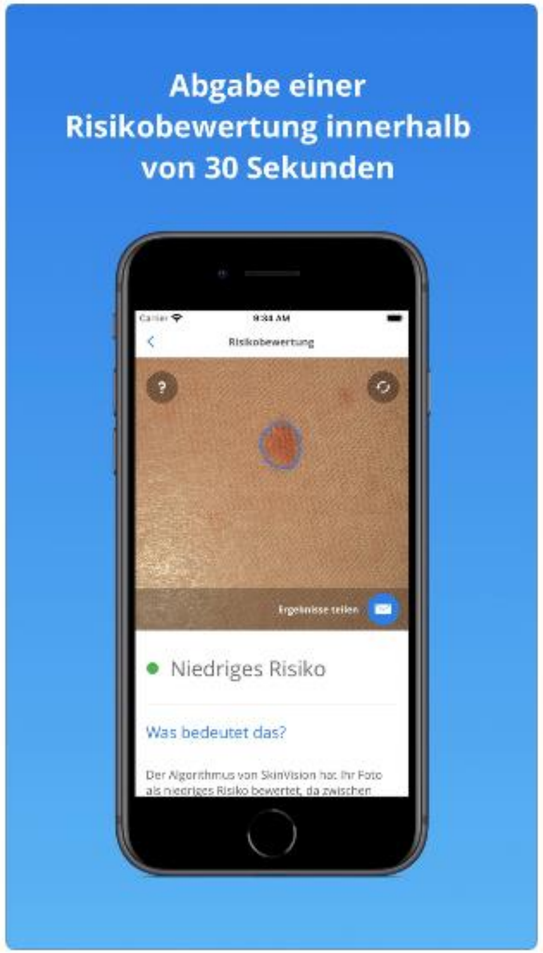
Classification of Benign and Malignant Melanomas

Incidence in Europe and North America 13 to 15 new illnesses in 100.000 residents. This results in a lifetime risk of a little more than one percent.

SkinVision: AI-powered app spots 95% of skin cancer!

Dermatologist-level classification of skin cancer with deep neural networks, Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun, Nature, volume, 542, 115–118 (2017)

SkinVision



Automatic External Defibrillator (AED)

In Germany, 100,000 to 200,000 people die every year from sudden ventricular fibrillation.

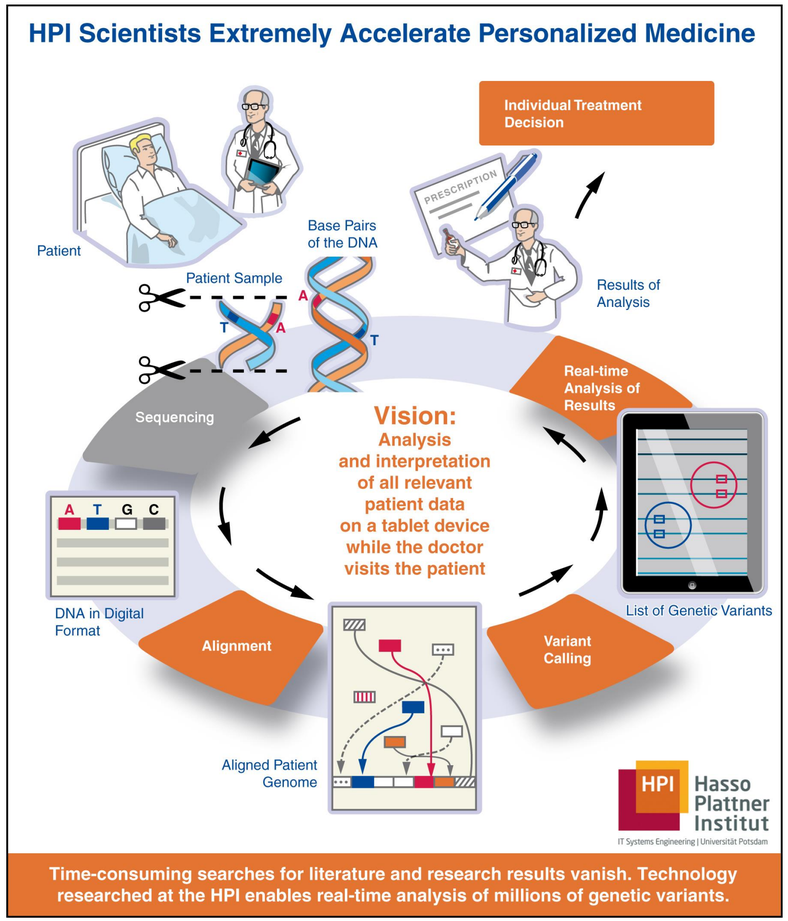
The automatic defibrillator must recognize whether it is ventricular fibrillation. Machine learning will be used for this in the next generation.

Journal of the American College of Cardiology. Volume 75, Issue 11 Supplement 1, March 2020. DOI: 10.1016/S0735-1097(20)34095-X, SPOTLIGHT ON SPECIAL TOPICS, DEVELOPMENT OF A CONVOLUTION NEURAL NETWORK FOR SHOCKABLE ARRHYTHMIA CLASSIFICATION WITHIN A NEXT GENERATION AUTOMATED EXTERNAL DEFIBRILLATOR
Christine Shen, et al.

Big Data for Health - Oncology

Transform data to knowledge → Providing knowledge for every physician

…. generates an evidence based hypothesis



3 Examples from the Institute of Biomedical Engineering

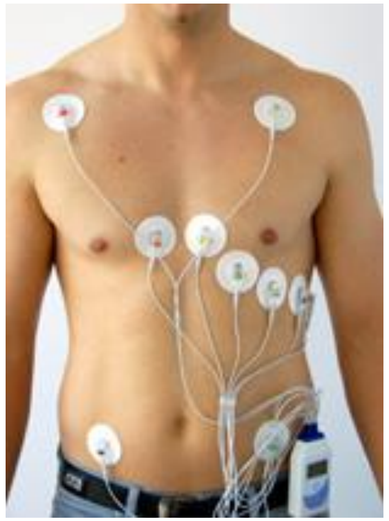
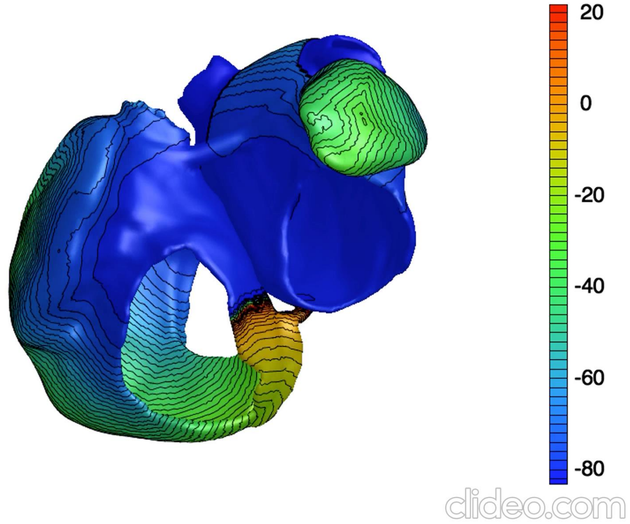
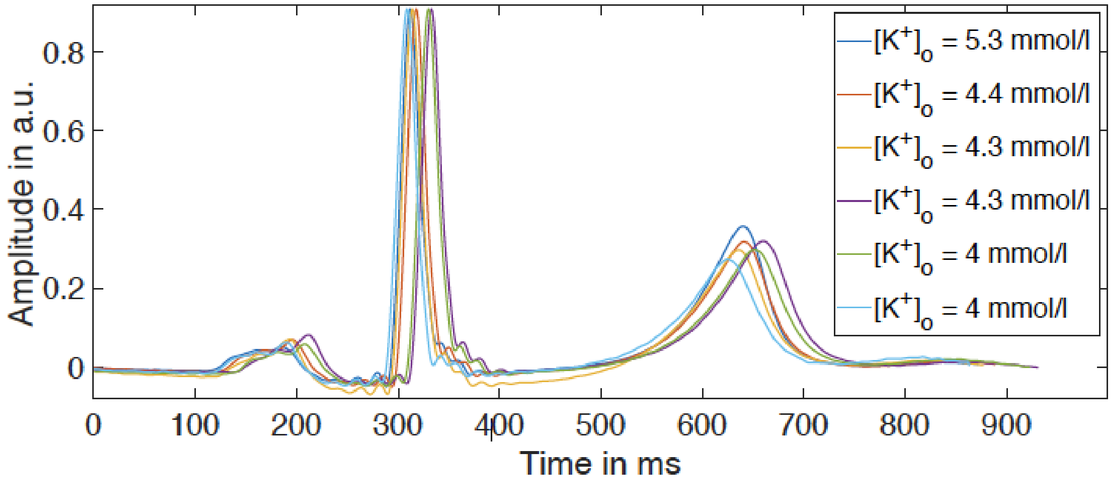


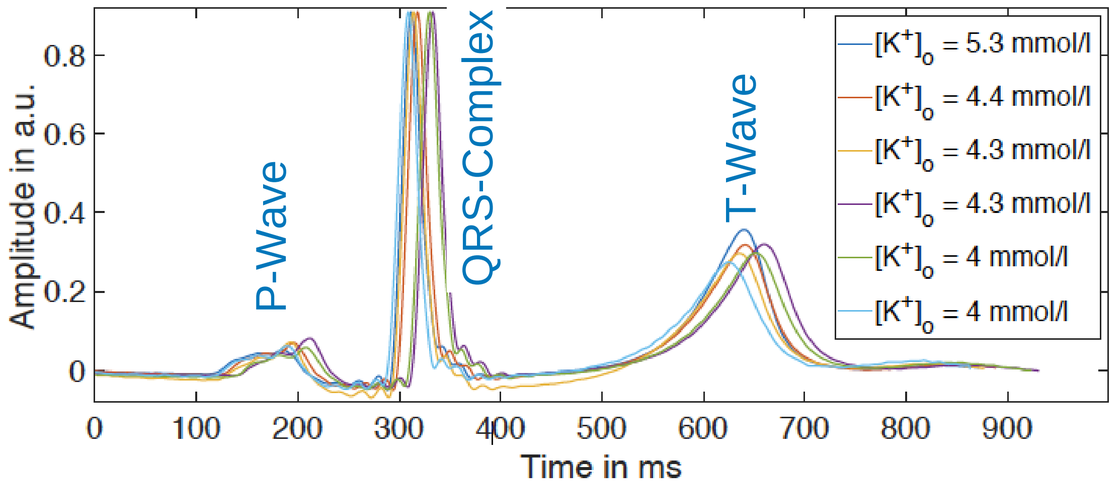
Foto: Corscience



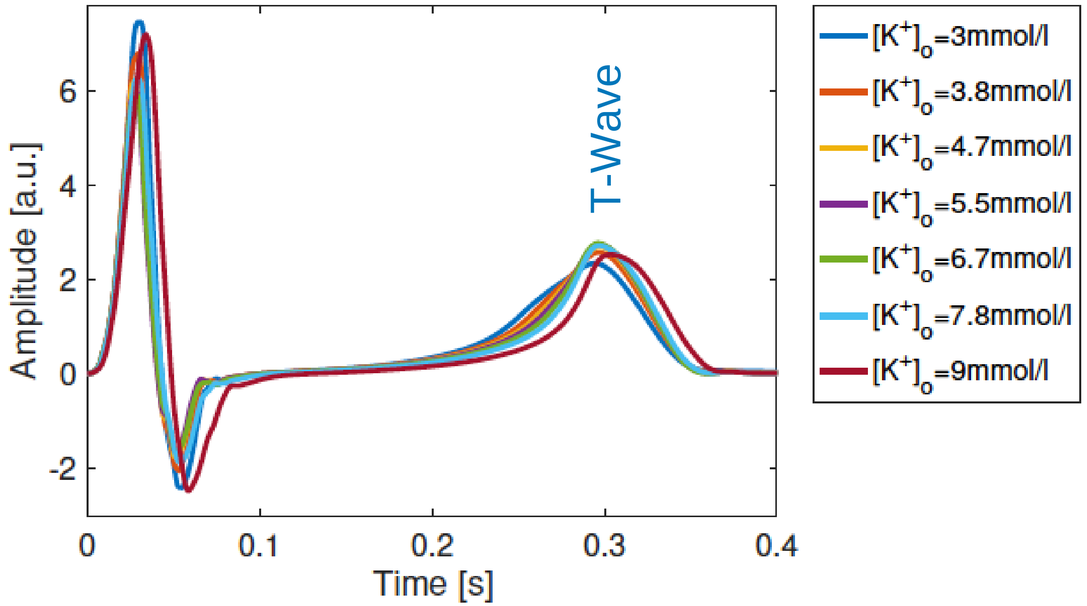


Estimating blood electrolyte concentration with the ECG

- Unbalanced electrolytes can lead to lethal arrhythmias.
- Patients suffering from kidney disease undergoing dialysis are in danger.
- Objective: detect dangerous unbalance of electrolytes using an ECG.
- Training of a Neural Net with simulated data and application to patient data

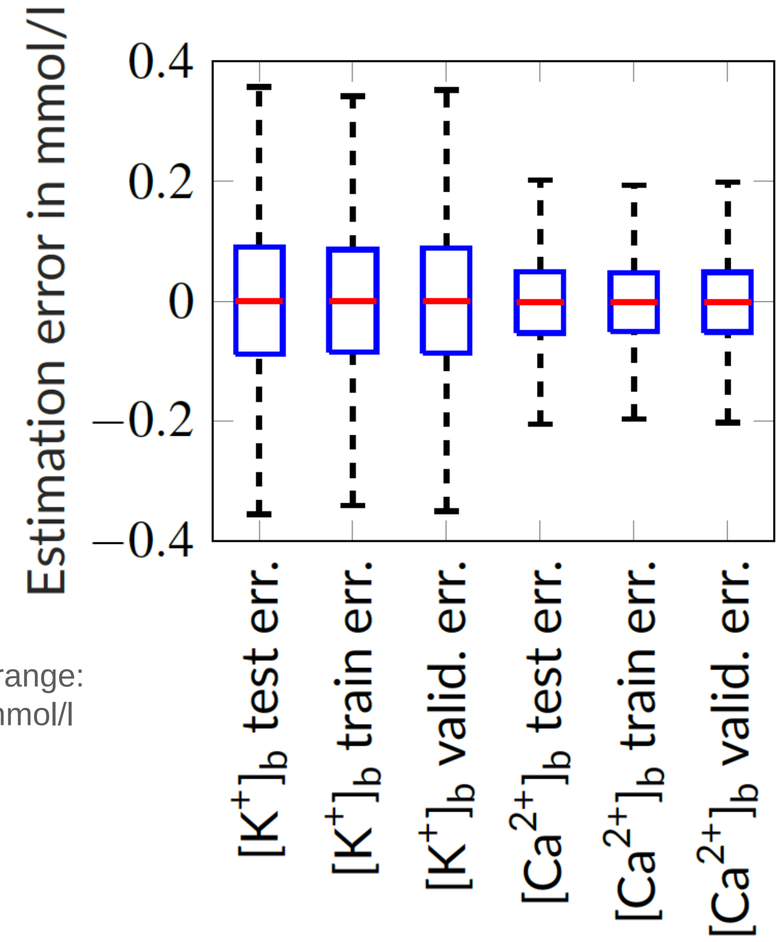


clinical ECG



simulated ECG

Estimating blood electrolyte concentration with the ECG



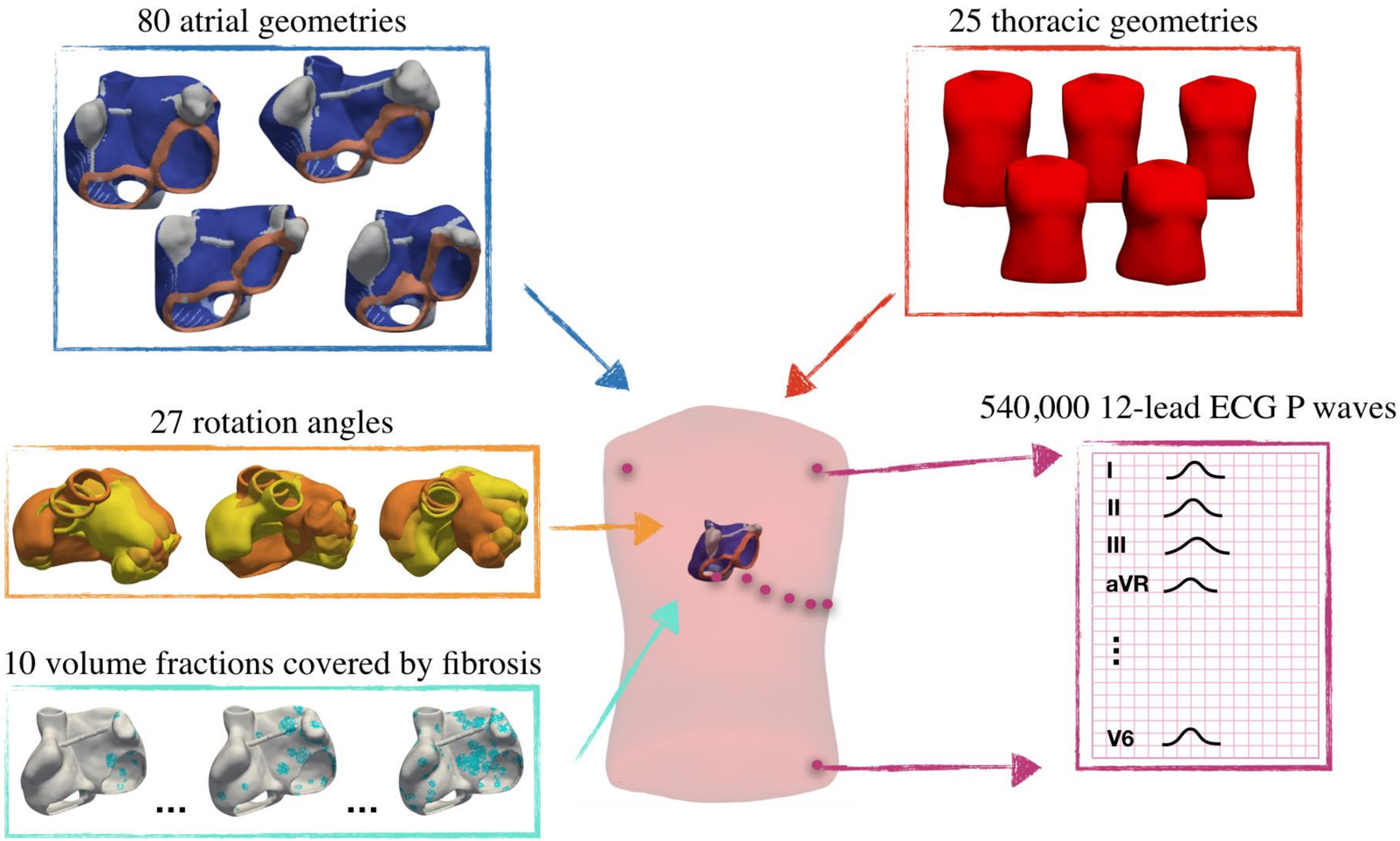
dynamic range:
3 bis 9 mmol/l

Regression with a neural network
(Regression with a polynomial fit)

ECG as a Tool to Estimate Potassium and Calcium Concentrations in the Extracellular Space, Pilia, N.; Dössel, O.; Lenis, G.; Loewe, A., CinC (2017) 44

ECG P-waves uncover the degree of fibrosis in the atria

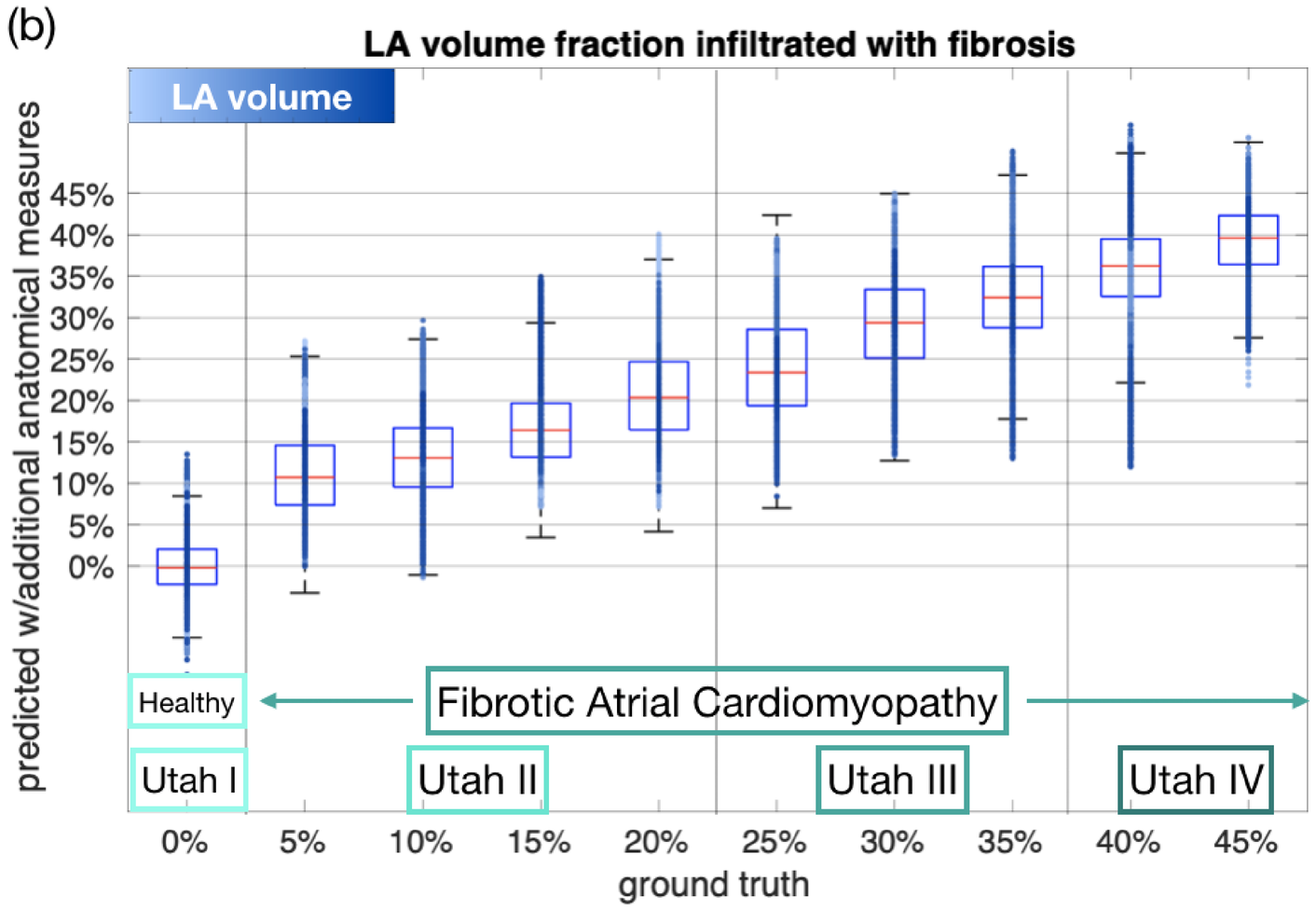
- Fibrosis in the atria is an indicator for susceptibility to Atrial Fibrillation.
- Quantifying the degree of fibrosis using an ECG can be an important biomarker.



generation of 540.000 P waves with known ground truth

Non-Invasive and Quantitative Estimation of Left Atrial Fibrosis Based on P Waves of the 12-Lead ECG—A Large-Scale Computational Study Covering Anatomical Variability Claudia Nagel et al. J. Clin. Med. 2021, 10, 1797. https://doi.org/10.3390/jcm10081797

ECG P-waves uncover the degree of fibrosis in the atria



Regression of the degree of fibrosis with a neural network.

18HLT07 MedalCare

Metrology of automated data analysis for cardiac arrhythmia management

Thursday, 3rd of November, 15:10
Sensitive Hearts: Challenges with Sensitivity Analysis of Cardiac Models
L. Wright, J. Venton

Classification of Atrial Flutter with the ECG

- Many patients suffer from atrial flutter. There are about 20 classes of AFlut.
- Knowing the type of flutter for a patient can speed up therapy.

Macroreentry: Tricuspid Valve

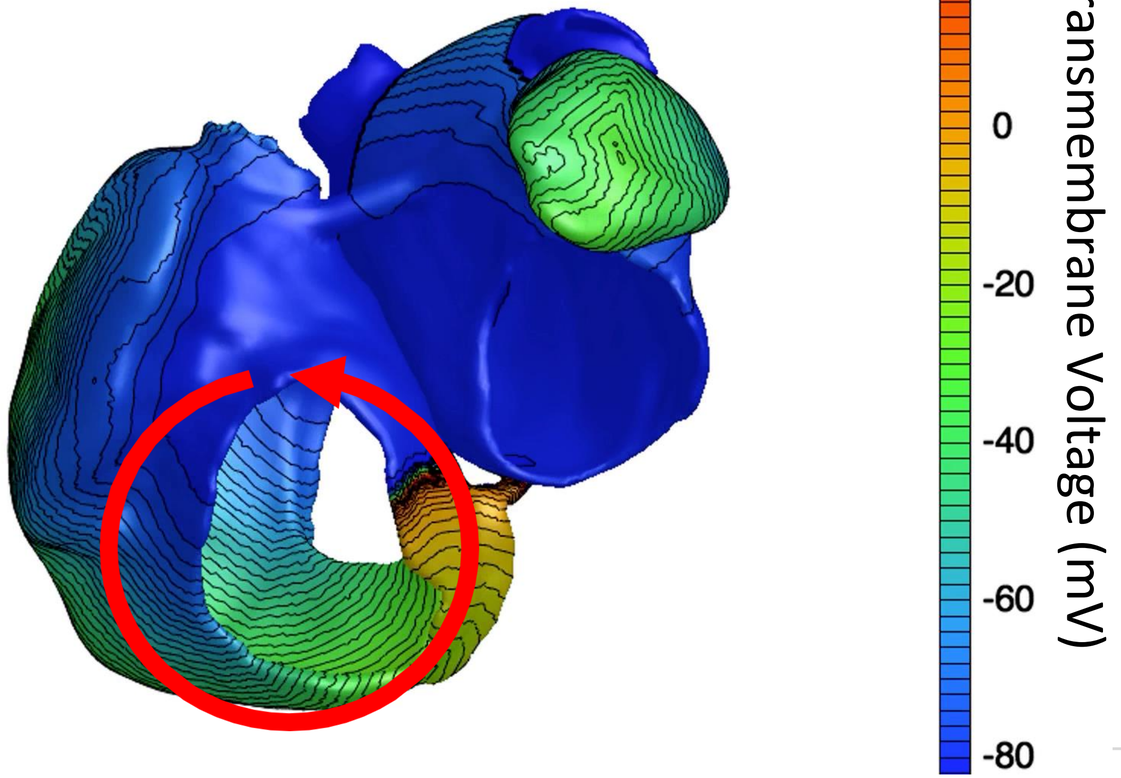
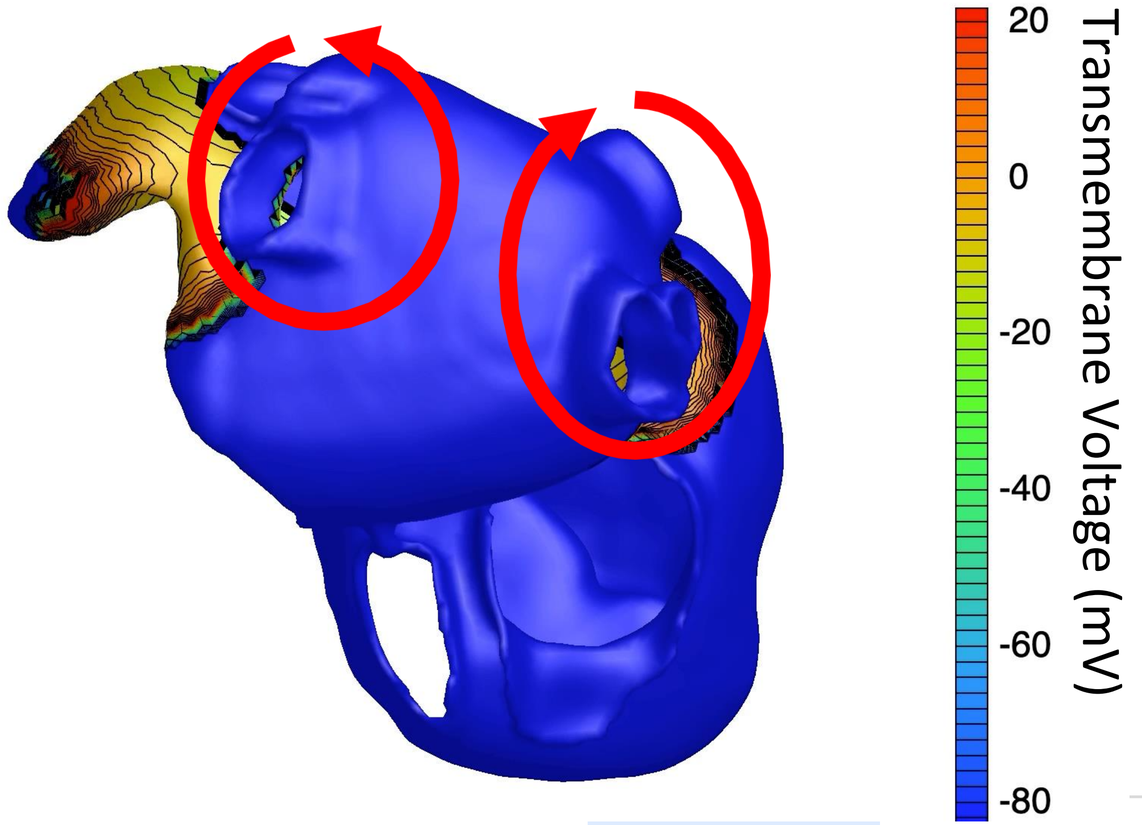


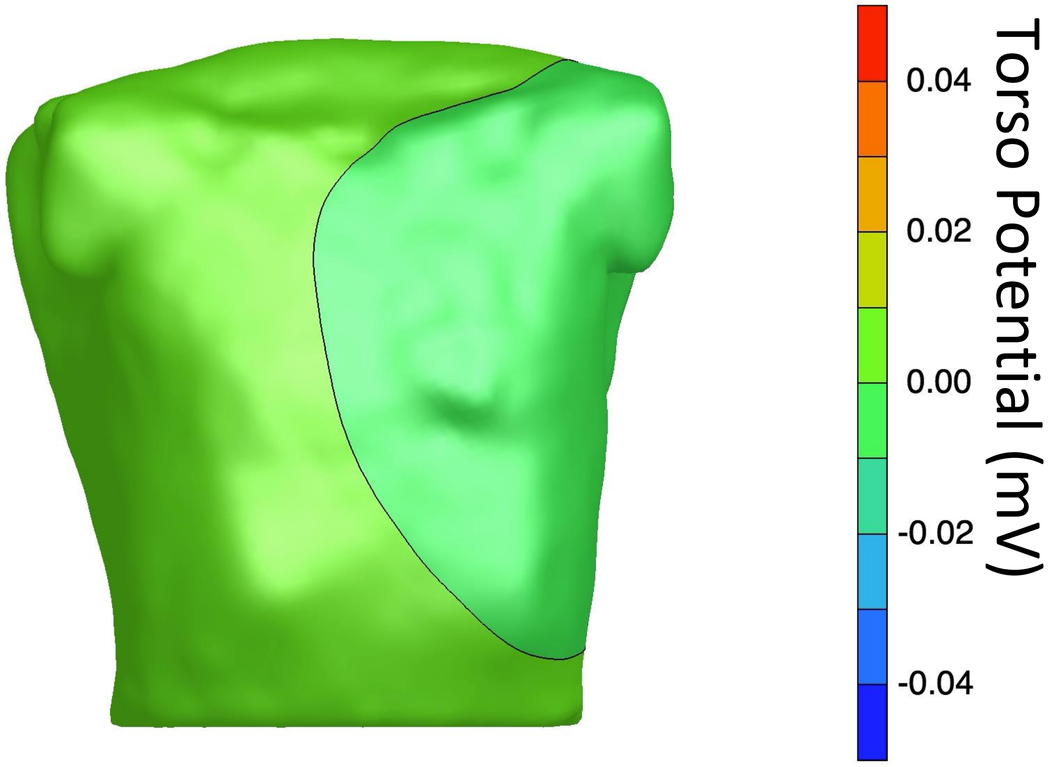
Figure-8 Macroreentry: ant PVs



Classification of Atrial Flutter with the ECG

Simulation of 1256 cases of Atrial Flutter and 12-lead ECG extraction

Body surface potential map



12-lead ECG (F-wave)

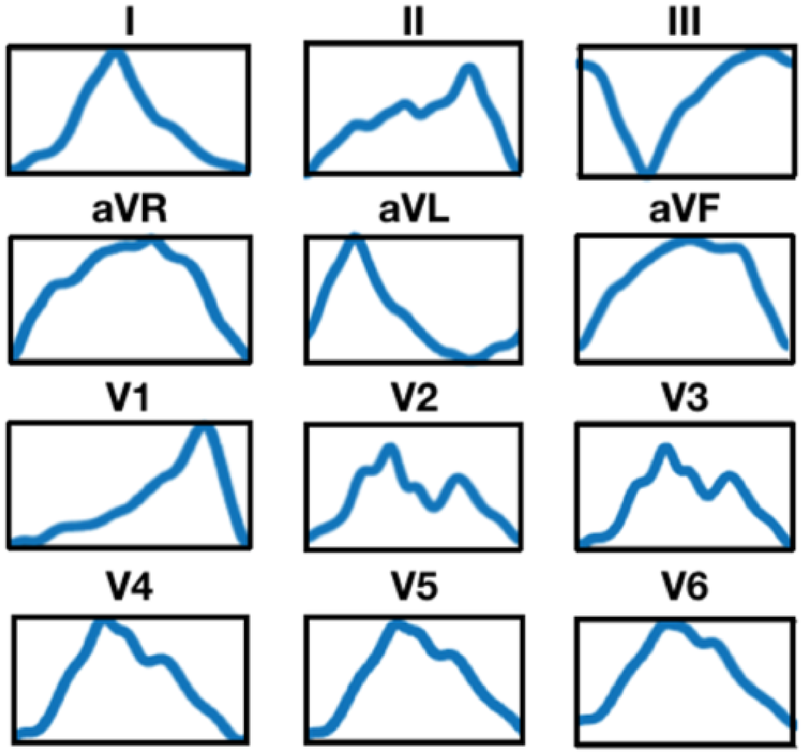


Figure-8 Macroreentry: ant PVs

Classification of Atrial Flutter with the ECG

Mechanism	Atrium	Position	Direction
Macroreentry	RA	Tricuspid Valve	ccw
Macroreentry	RA	Tricuspid Valve	cw
Macroreentry	LA	Mitral Valve	ccw
Macroreentry	LA	Mitral Valve	cw
Scar-related Reentry	LA	LPV	post
Scar-related Reentry	LA	LPV	ant
Scar-related Reentry	LA	RPV	post
Scar-related Reentry	LA	RPV	ant
Figure-8 Macroreentry	LA	Both PVs	ant
Figure-8 Macroreentry	LA	Both PVs	post
Figure-8 Macroreentry	LA	RPVs	ant
Focal Source	LA	RSPV anterior	
Focal Source	LA	RSPV posterior	
Focal Source	LA	LSPV anterior	
Focal Source	LA	LSPV posterior	
Microreentry	LA	ant MV annulus	
Microreentry	LA	ant LAA	
Microreentry	LA	ant RSPV	
Figure-8 Microreentry	LA	ant	
Microreentry	LA	post wall	

- 151 extracted features
- Greedy forward feature selection
- Random dataset division (70% training, 15% validation, 15% test set)
- Radial basis neural network (rbNN)

- Hit rate around 60%

Non-Invasive Characterization of Atrial Flutter Mechanisms Using Recurrence Quantification Analysis on the ECG: A Computational Study, Luongo, Giorgio; Schuler, Steffen; Luik, Armin; Almeida, Tiago P.; Soriano, Diogo C.; Dossel, Olaf; Loewe, Axel, IEEE Trans Biomed Eng (2021 Jan 1) 68 (3): 914-925.

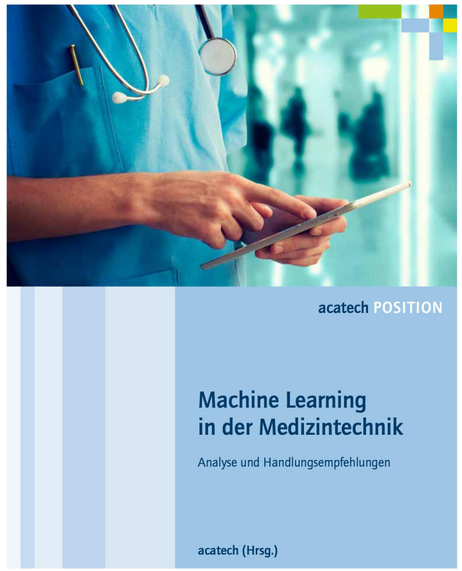
Classification Scheme for ML in Medical Technology

ML Applications Group 1:
Applications for medical lay people without a medical doctor involved (often not a medical device)

ML Applications Group 2:
Classification of patients into groups of different priority (Triage, Organ Transplant)

ML Applications Group 3:
Support of medical doctors in diagnosis and therapy

ML Applications Group 4:
Autonomous medical devices - no medical doctor involved

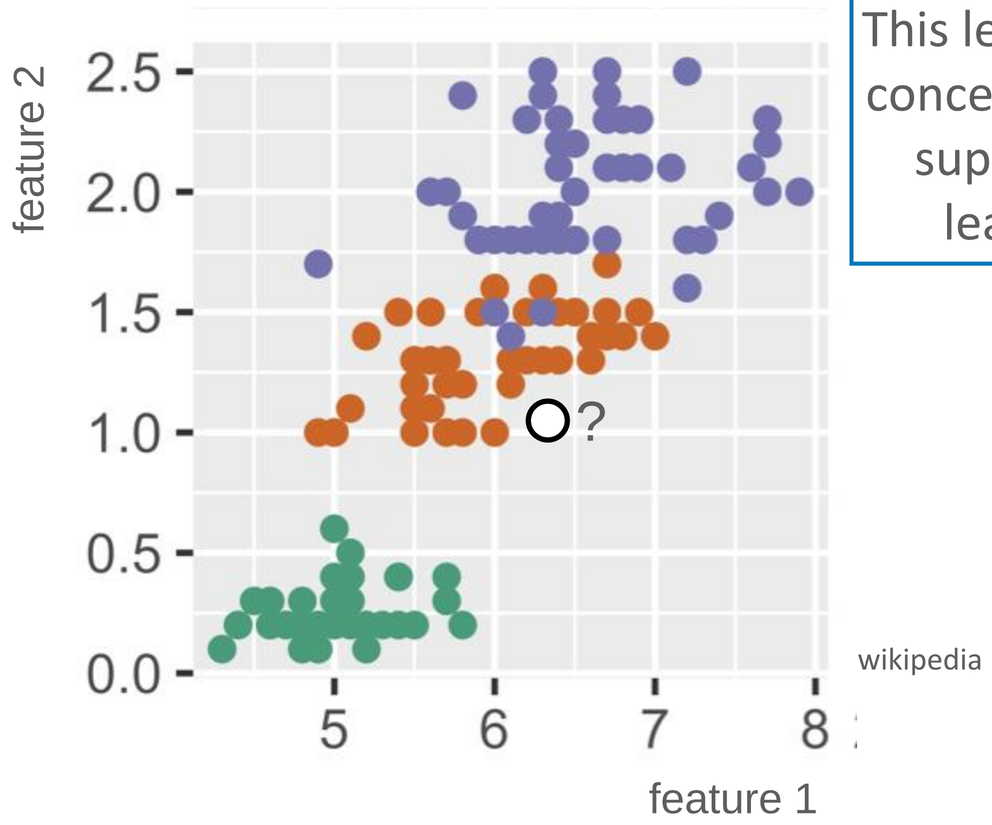


ML Option:
The algorithm is continuously learning and adapts itself to the individual patient

Artificial Intelligence and Machine Learning in Medicine

- Some Examples and 3 Projects at IBT
- Some Basics of ML
- How can we Measure the Quality of an AI / ML Algorithm
- Ethical Aspects
- Regulatory Aspects
- More Data for Research - Data Protection and Reference Data
- Some Statements

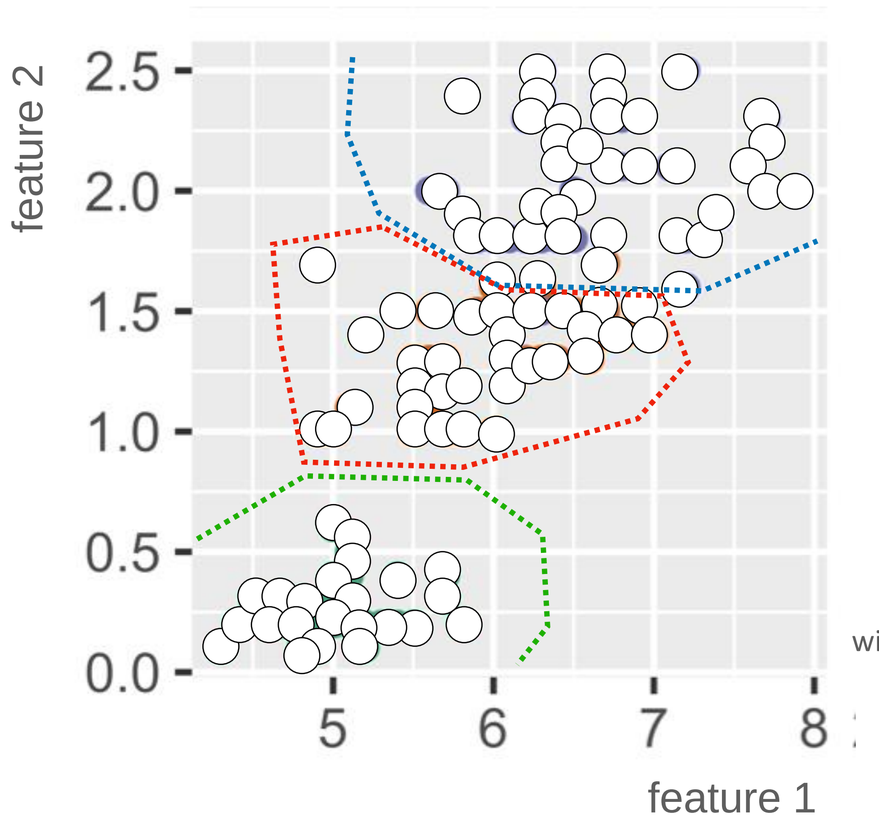
Supervised and Unsupervised Learning



wikipedia

This lecture will concentrate on supervised learning

In supervised learning we have a data set with known ground truth.



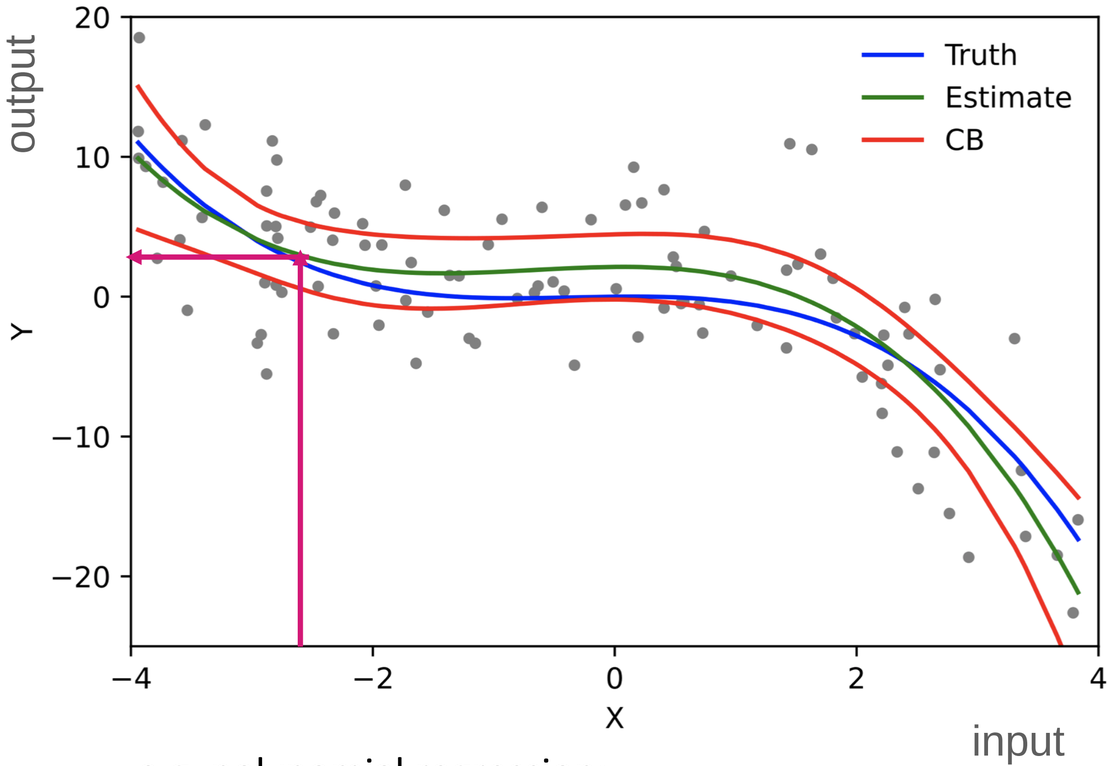
wikipedia

first define a distance

In unsupervised learning we ask the ML algorithm to suggest a meaningful subdivision into classes.

Regression and Classification

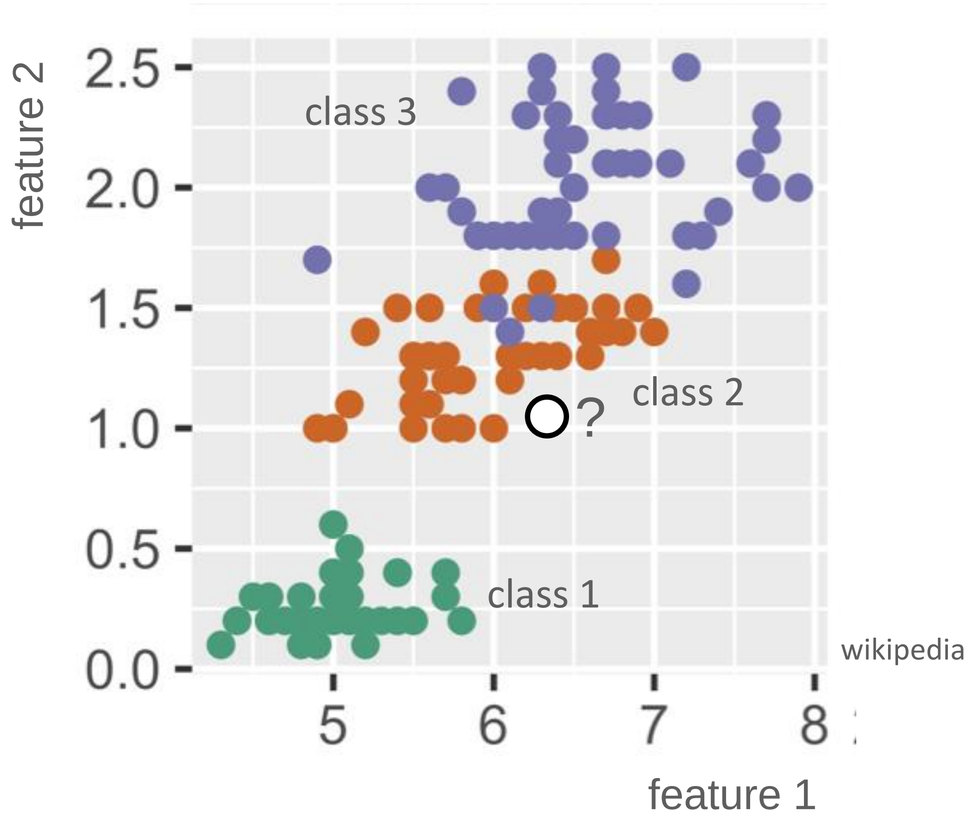
Regression delivers a **number**.



wikipedia

<< logistic regression >>

Classification delivers a **class.**



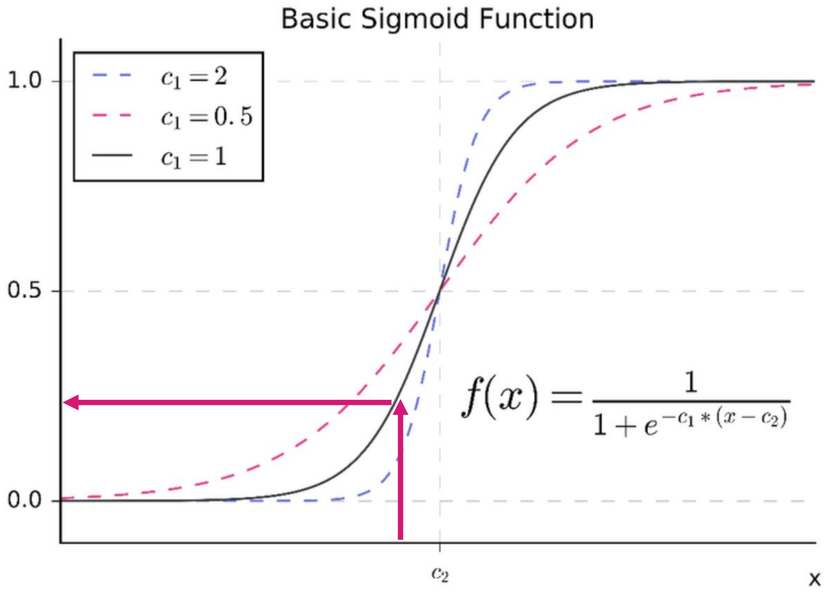
wikipedia

e.g. polynomial regression:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon.$$

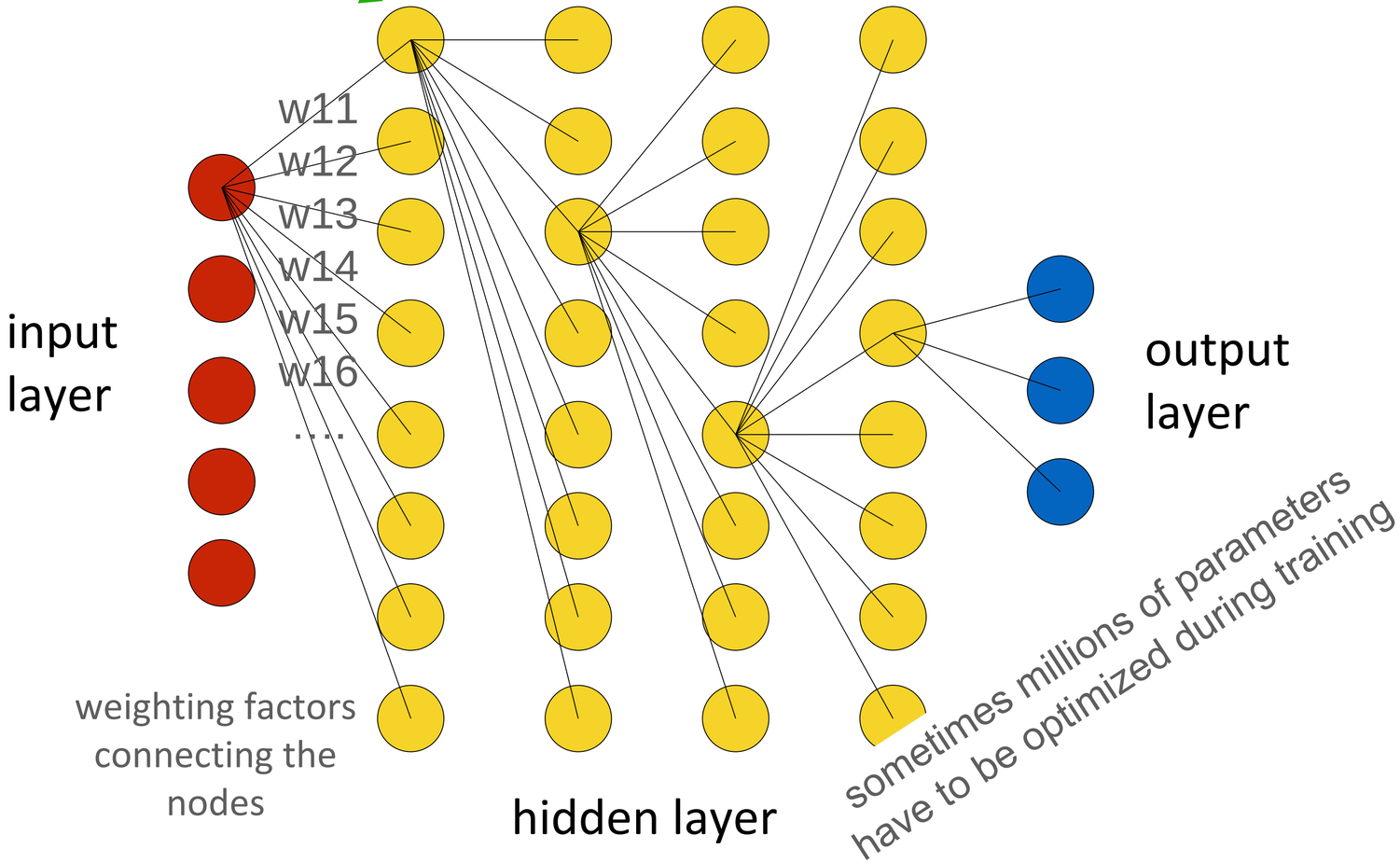
>> multivariate polynomial regression

Neural Networks and Deep Neural Networks



non-linear function in each node (bias)

wikipedia



Optimization needs a cost function / loss function!

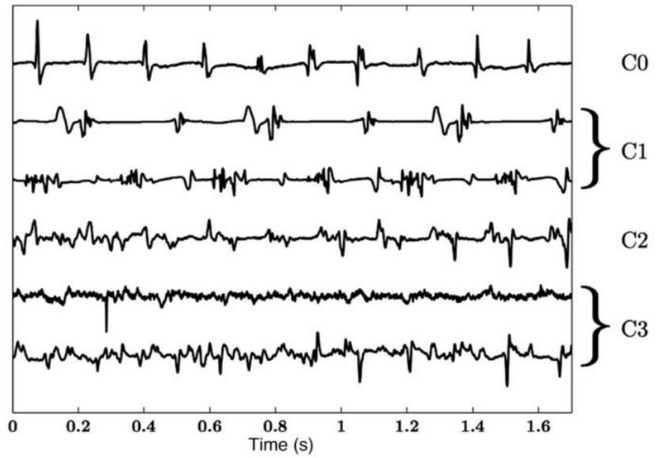
- Regression problems
  - root mean squared error / median
  - cross correlation / mutual information
  - Jaccard index / Dice Coefficient / Hausdorff metric
- Classification problems
  - hit rate ???
  - sensitivity & specificity
  - accuracy and F1 score

see more in the chapter on Quality Measures

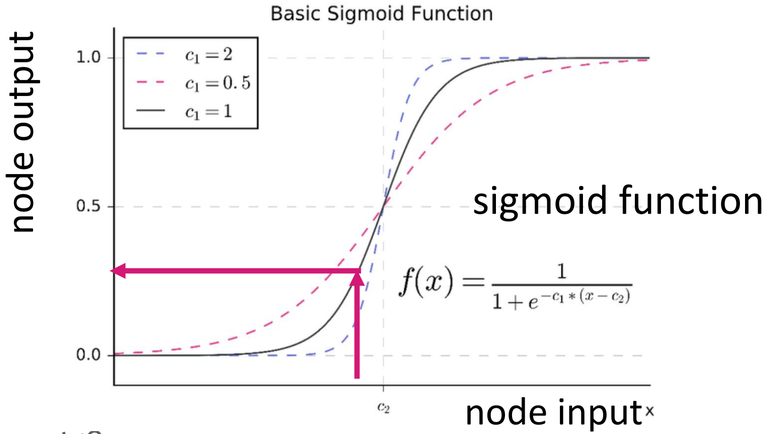
The Zoo of Neural Network AI and ML

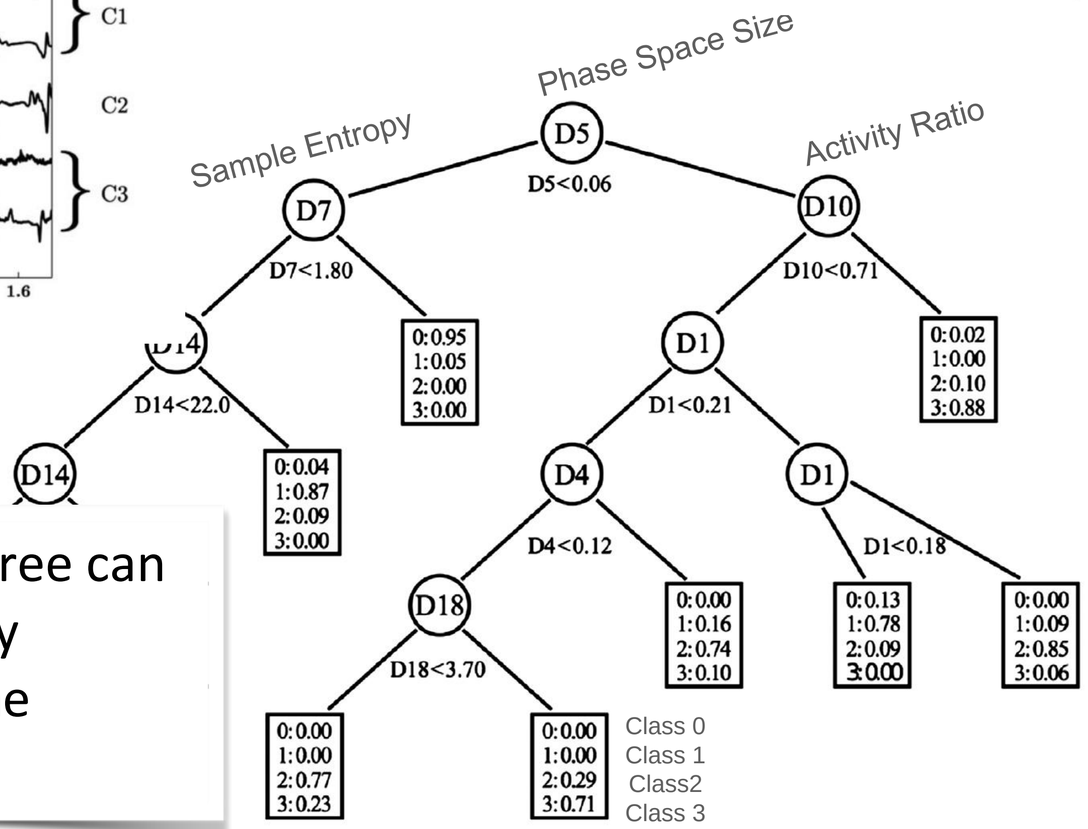
- Fully Connected Neural Networks
- Convolutional Neural Networks (CNN)
  - ResNet
  - LeNet
  - AlexNet
  - VGG
  - GoogLeNet
  - U-Net
- Recurrent Neural Networks (RNN)
  - LSTM (Long short-term memory)
  - GRU
  - ESN
- Transformer
  - BERT
  - GPT
- Spiking Neural Networks

Decision Tree and Fuzzy (Probabilistic) Decision Tree



4 classes of intracardiac electrograms





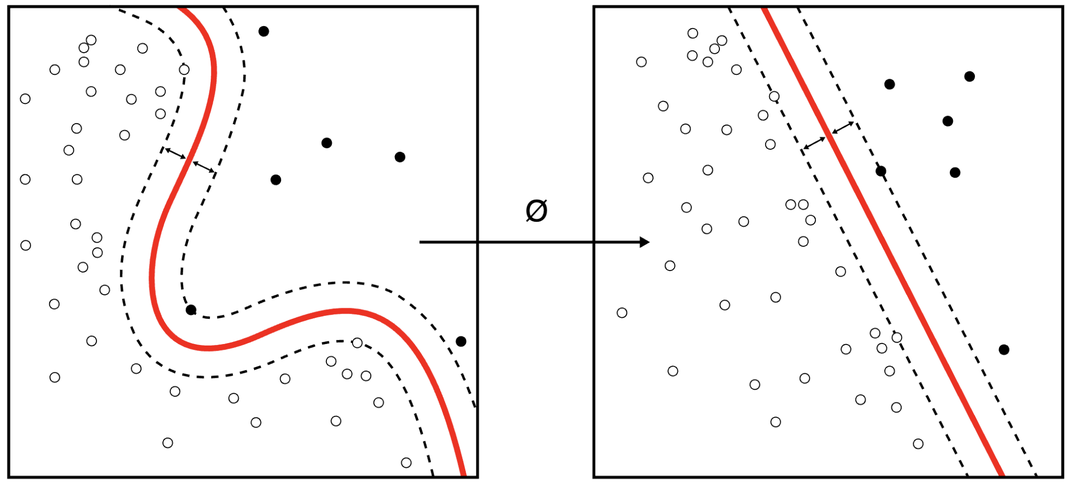
The fuzzy decision tree can show the most likely classification and the second best.

At each node the chosen descriptor and its split value is depicted. For each leaf node the membership result is shown. The correct rate for this tree is 86.1%.

Fuzzy decision tree to classify complex fractionated atrial electrograms
Schilling, C.; Keller, M.; Scherr, D.; Oesterlein, T.; Haissaguerre, M.; Schmitt, C.; Dössel, O.; Luik, A.
Biomed Tech (Berl) (2015 Jan 1) 60 (3): 245-255.

Support Vector Machine

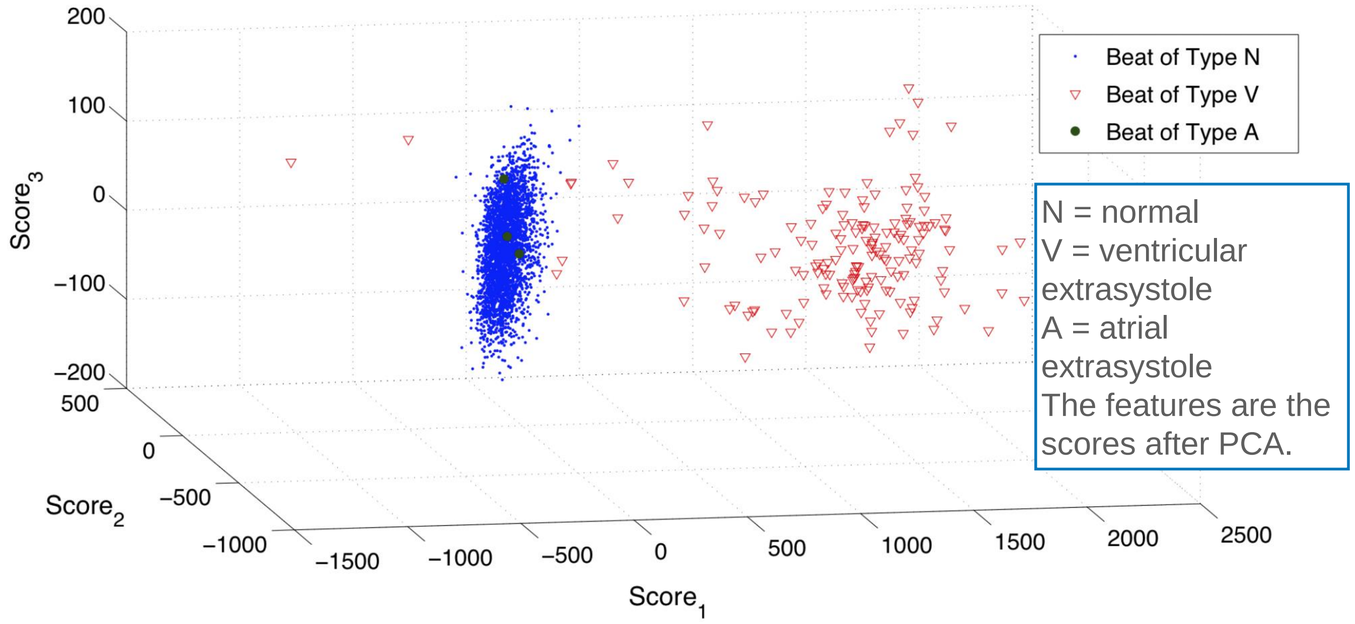
In support-vector machines, a data point is viewed as a p-dimensional vector and we want to separate such points with a (p−1)-dimensional hyperplane in terms of a linear classifier.



wikipedia

By Alisneaky, svg version by User:Zirguezi

Automatic detection and classification of ectopic beats in the ECG

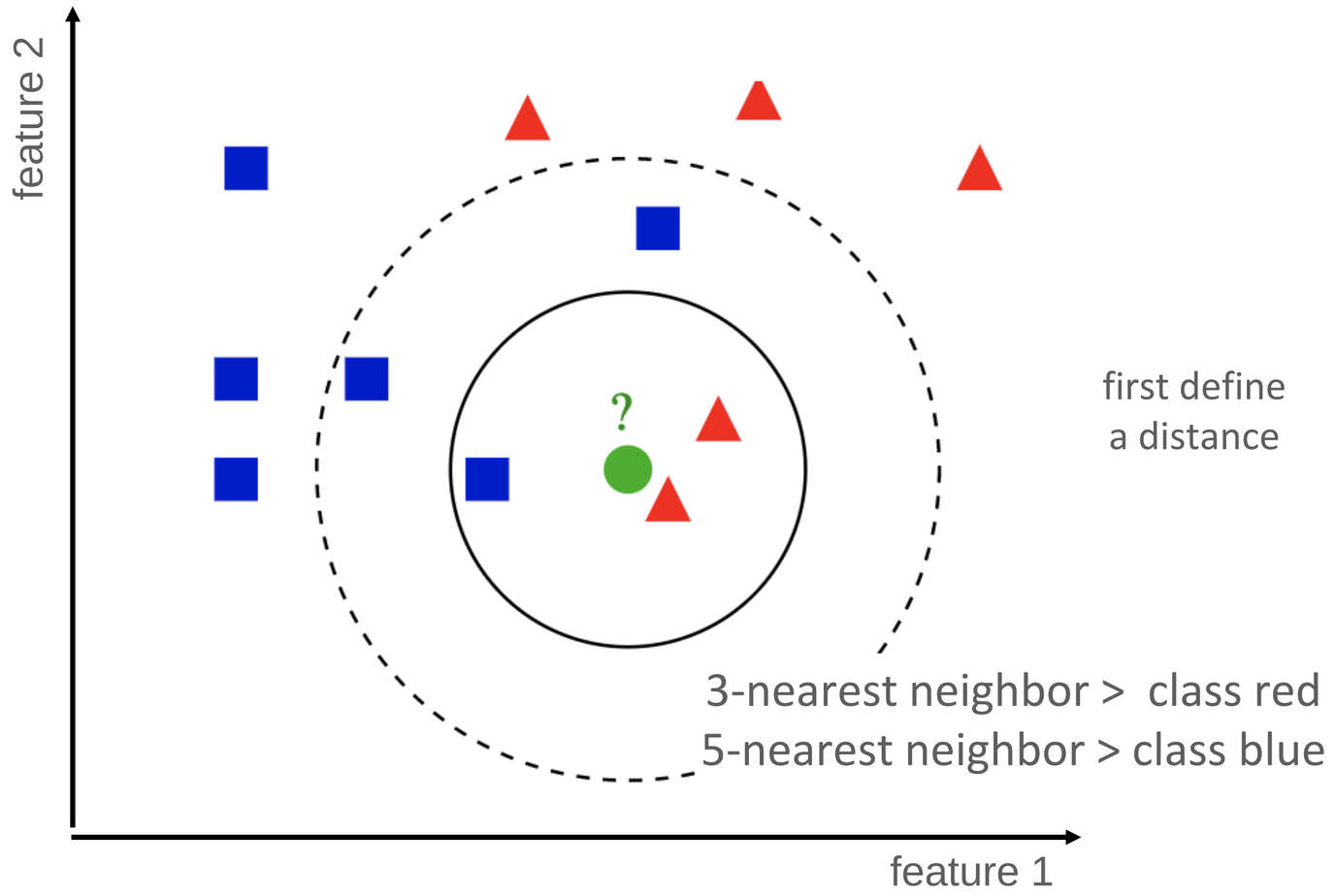


N = normal
V = ventricular extrasystole
A = atrial extrasystole
The features are the scores after PCA.

Automatic detection and classification of ectopic beats in the ECG using a Support Vector Machine
Lenis, G.; Baas, T.; Dössel, O.
BMT2011 (2011 Jan 1)

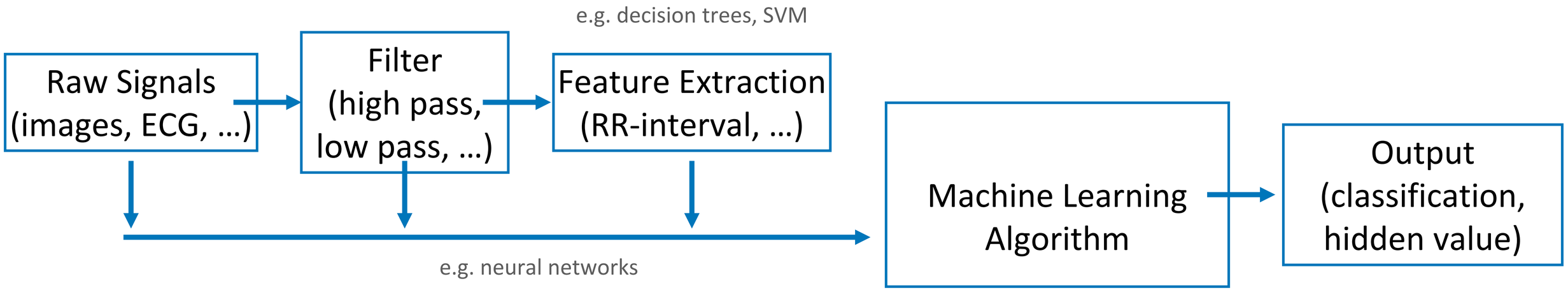
k-Nearest Neighbor

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small).



wikipedia

Raw Signals or Filtered Signals or Features ?



There are various pathways through this system. The pathway that was chosen for training must be identical to the pathway that is finally used for output generation.

18HLT07 MedalCare

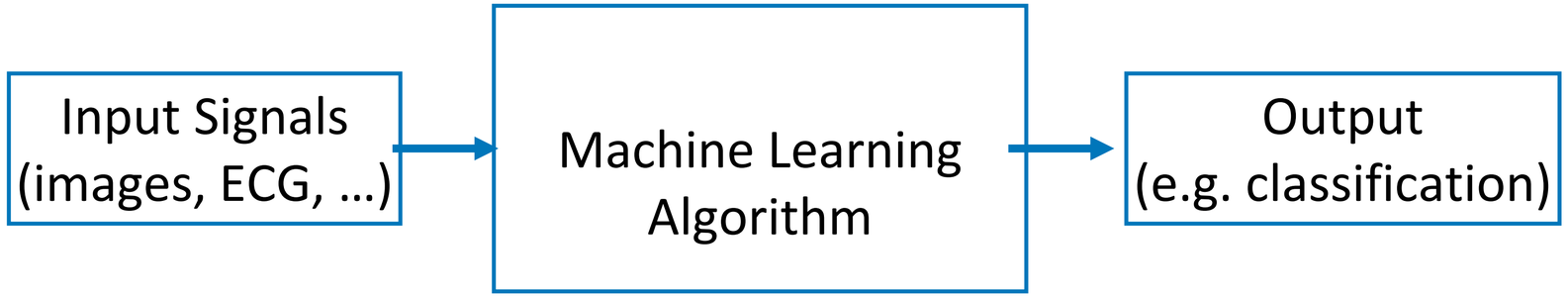
Metrology of automated data analysis for cardiac arrhythmia management

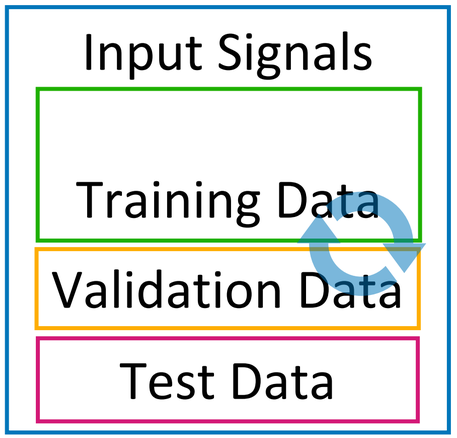
Artificial Intelligence and Machine Learning in Medicine

- Some Examples and 4 Projects at IBT
- Some Basics of ML
- How can we Measure the Quality of an AI / ML Algorithm
- Ethical Aspects
- Regulatory Aspects
- More Data for Research - Data Protection and Reference Data
- Some Statements

Prerequisite: Clear Separation of Training Dataset - Validation Dataset - Test-Dataset

…to reduce the problem of overfitting and loss of generalization





used for training

used for optimizing the training procedure

used for final quality check
(sensitivity, specificity, accuracy, …)

another option:
n-fold cross validation

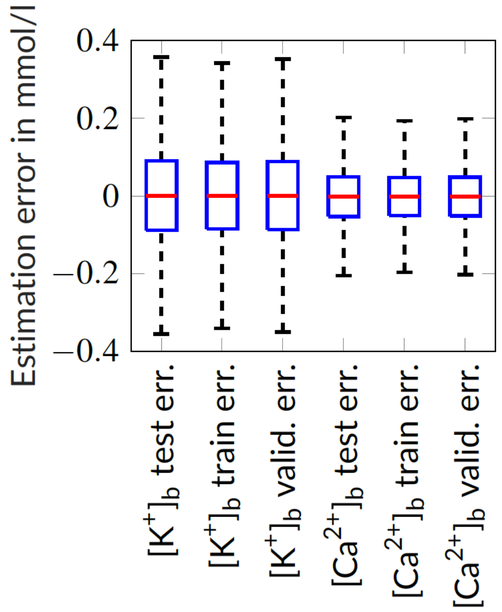
leave-one-out cross validation

Better: independent reference data, hidden to the developer!

How to Measure Prediction Error of a Regression

compare to the chapter on loss function

- Regression (0D)
  - Root Mean Squared Error
  - ….



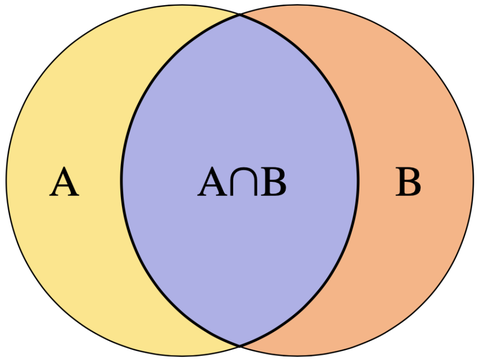
- Regression (1D)
  - Root mean squared error
  - Cross Correlation
  - Mutual Information
  - …….

$$R_{xy}(t_1, t_2) = E\{\mathbf{X}(t_1) \cdot \mathbf{Y}(t_2)\}$$

$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_{(X,Y)}(x,y) \log\left(\frac{p_{(X,Y)}(x,y)}{p_X(x)\, p_Y(y)}\right)$$

- Regression (2D)
  - Jaccard Index
  - Hausdorff Distance
  - Dice Coefficient ….

e.g. segmentation



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

How to Measure Prediction Accuracy of a Classification

- Confusion Matrix

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection

wikipedia

sensitivity:
I do not want to overlook a single TP

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

specificity, selectivity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}$$

specificity:
how well can I identify TNs ?

Observe the Receiver Operating Curve (ROC)!

More Information from the Confusion Matrix

		Predicted condition			
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ra = FN
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnosti =
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{TP}{TP + FN + FP}$

... always problems with unbalanced data

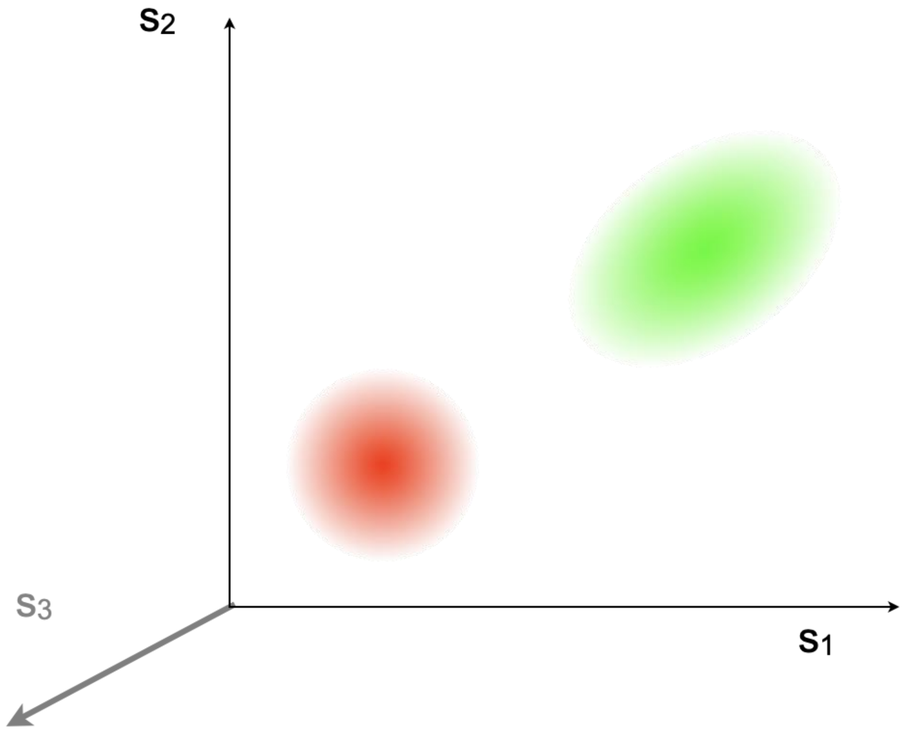
wikipedia

Uncertainty Quantification in ML

- Epistemic uncertainty
  - reducable with more information, e.g. due to model uncertainty
- Aleatoric uncertainty
  - irreducable, e.g. due to noise

- Global uncertainty
  - valid for the cohort of all patients
- Local uncertainty
  - valid for the patient under test

- Model agnostic methods
  - valid for all ML methods
- Model specific methods
  - working only with one method of ML



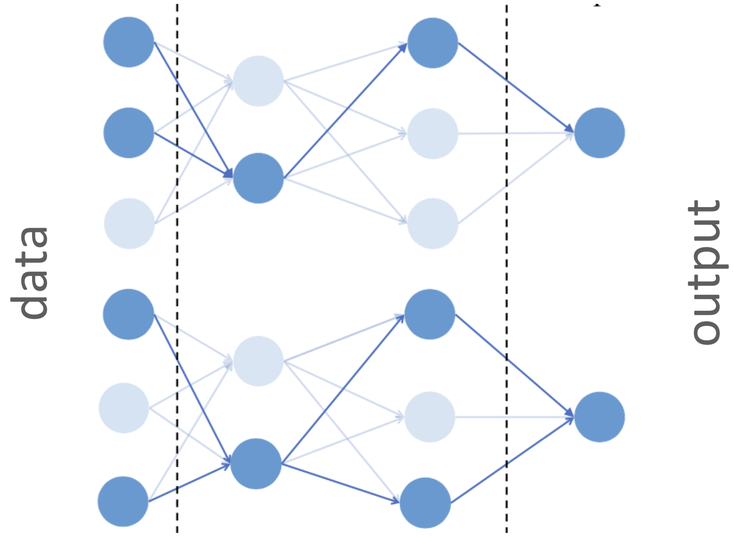
s_1 and s_2 : vectors in state space

uncertainty of state B

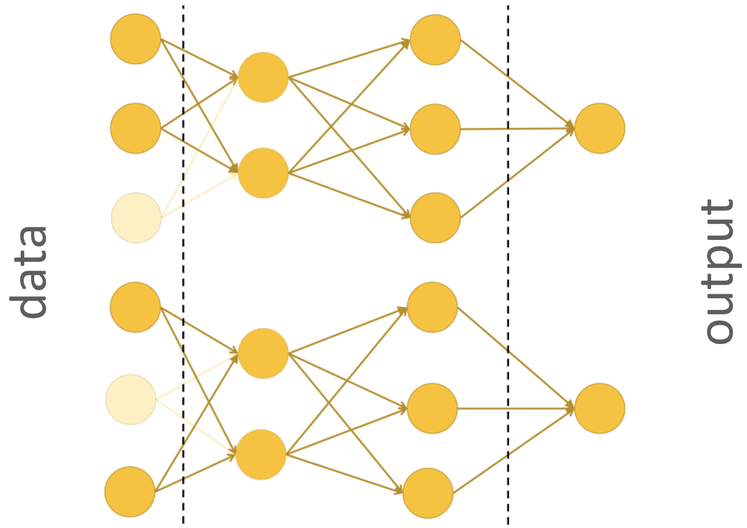
uncertainty of state A

s_3 uncertainty by omitting one dimension in state space

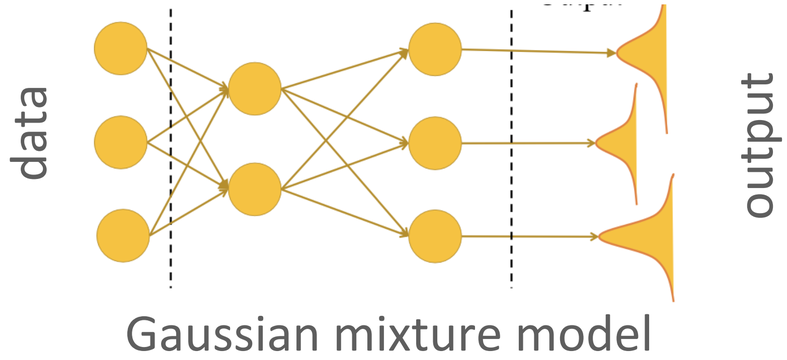
Uncertainty Quantification in Deep Learning



Monte Carlo dropout



bootstrap method



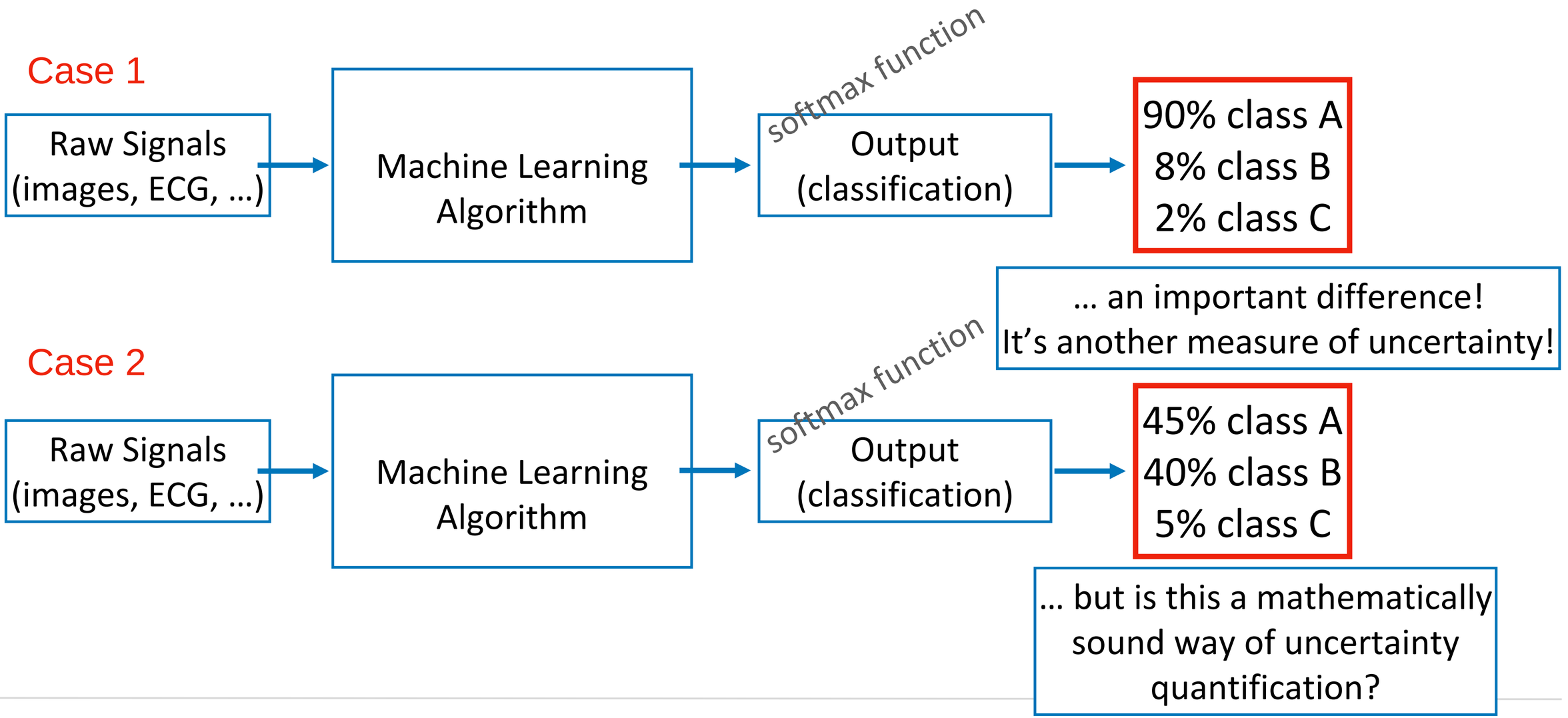
Gaussian mixture model

C. Hubschneider, R. Hutmacher, J.M. Zöllner, Calibrating uncertainty models for steering angle estimation, in: IEEE Intelligent Transportation Systems Conference, 2022

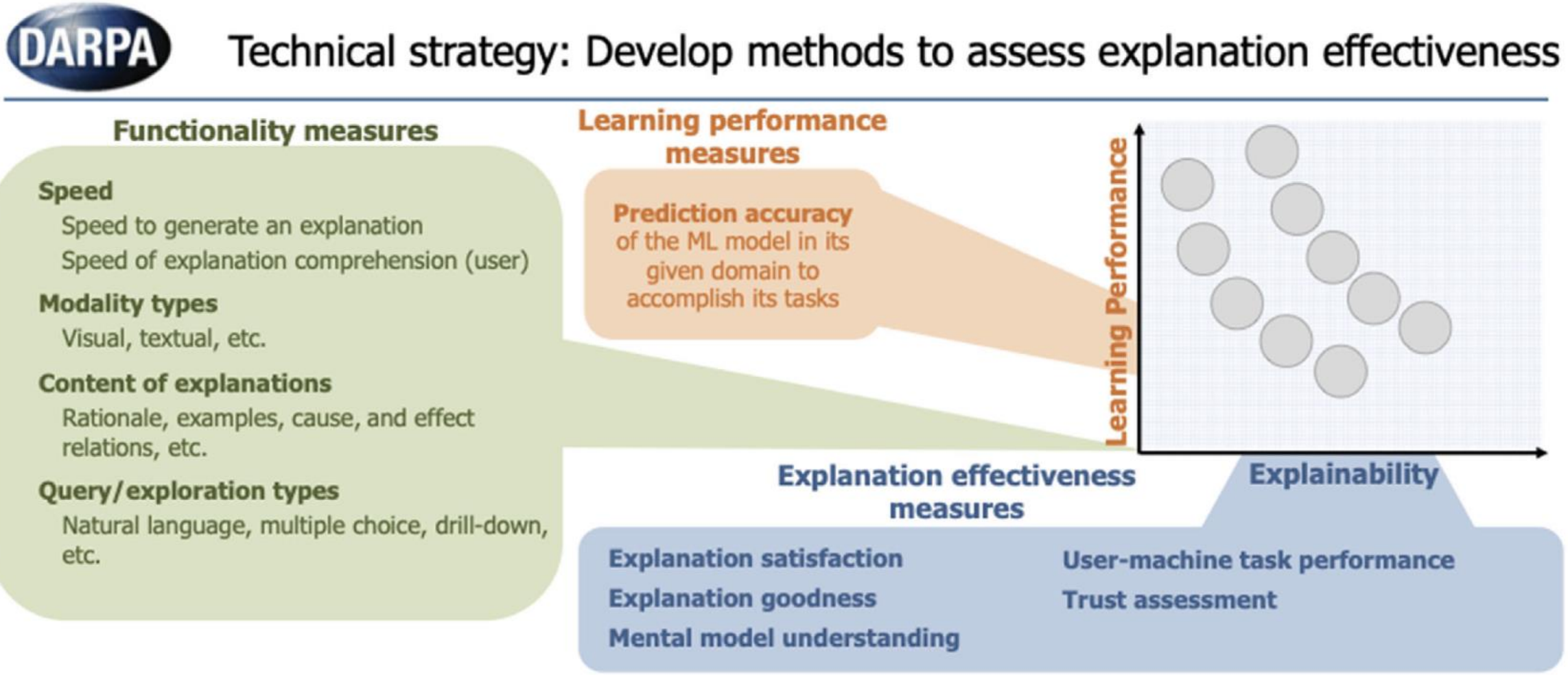
736 references

A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion 76 (2021) 243–297
Moloud Abdar et al.

The Preferred Classification and the Second and Third Best Classification



DARPA's explainable AI (XAI) program



https://onlinelibrary.wiley.com/doi/epdf/10.1002/ail2.61

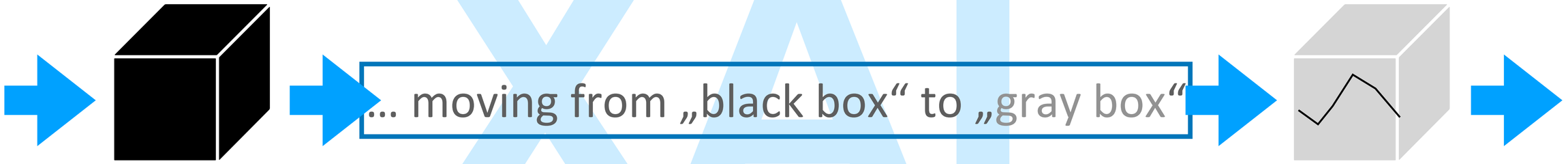
Explainable Artificial Intelligence XAI

label: „ML inside“
data are open
algorithm is open

transparent

explainable

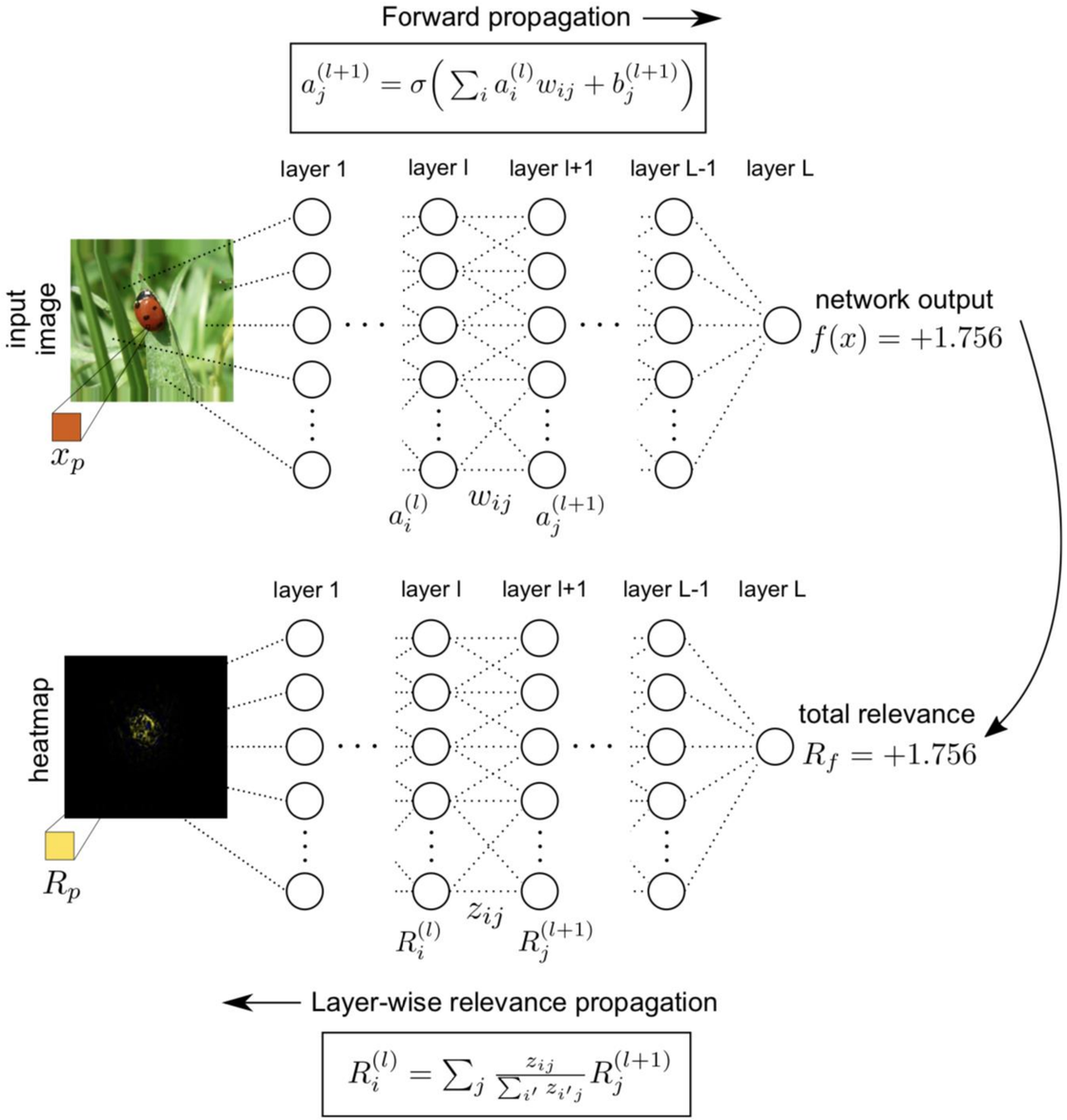
interpretable
comprehensible
understandable
cause-and-effect logic



Explainable AI - XAI

- Layerwise Relevance Propagation

Layerwise Relevance Propagation (LRP),
Counterfactual Method,
Local Interpretable Model-agnostic Explanation (LIME),
Generalized Additive Model (GAM),
Reversed Time Attention Model (RETAIN),
Black Box Explanations Through Transparent Approximations (BETA),
Deep Taylor Decomposition,
Bayesian Rule Lists (BRL)

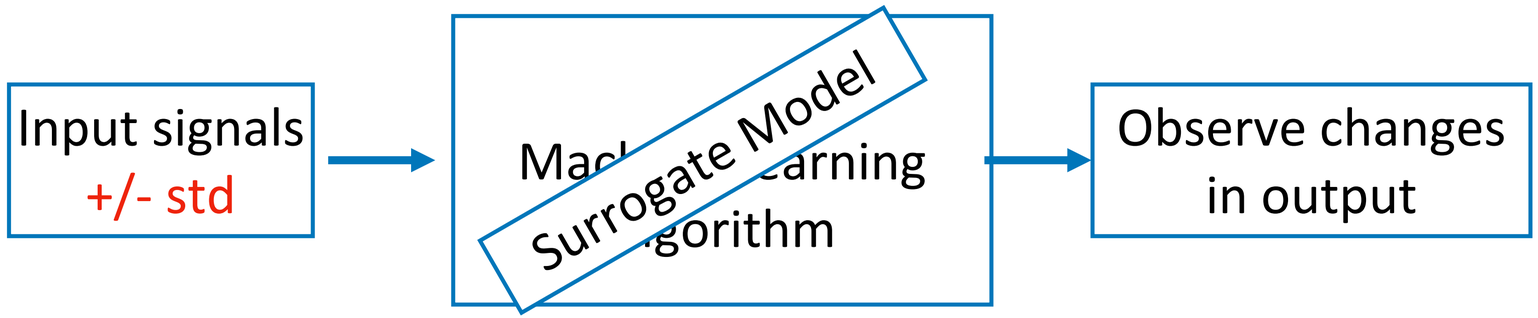


Moving from the „Black Box“ to the „Gray Box“

Layer-wise Relevance Propagation for Deep Neural Network Architectures
Alexander Binder, Sebastian Bach, Gregoire Montavon, Klaus-Robert Müller, and Wojciech Samek

Sensitivity to Input Data Variation

- Which input data have the strongest influence on the result?
- Is the result robust to small errors in the input data?



- Monte Carlo
- Surrogate Model and Polynomial Chaos >> Sobol Indices

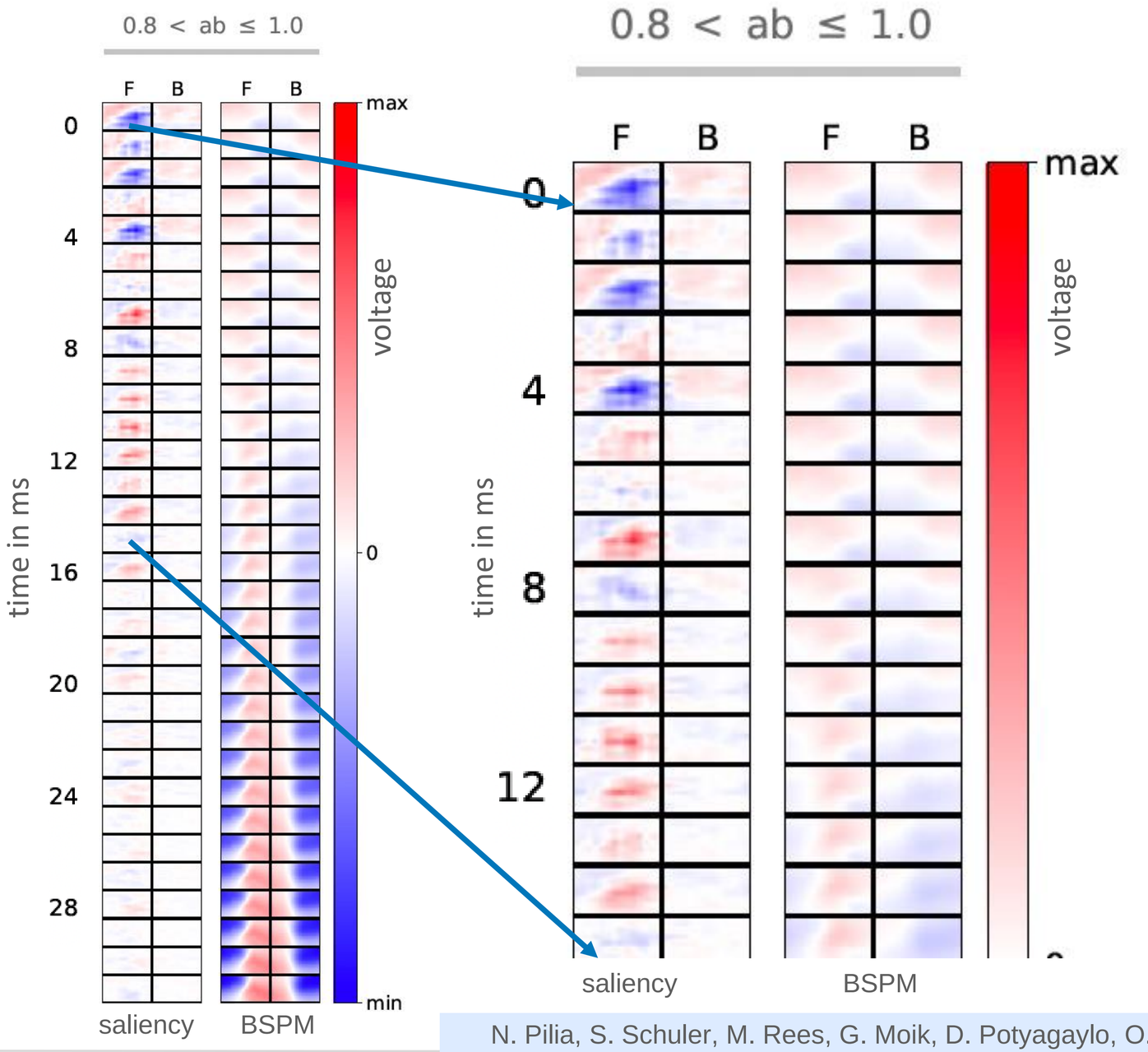
XAI

18HLT07 MedalCare

Metrology of automated data analysis for cardiac arrhythmia management

Saliency Maps
… for finding the origin of a ventricular extrasystole

depicted are the first 32ms,
left: saliency map
right: Body Surface Potential Map
F: Front
B: Back



K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps.” CoRR, vol. abs/1312.6034, 2013.

N. Pilia, S. Schuler, M. Rees, G. Moik, D. Potyagaylo, O. Dössel, and A. Loewe, Non-invasive Localization of the Ventricular Excitation Origin Without Patient-specific Geometries Using Deep Learning. In eprint, 2022

Ranking of Feature Importance for ECG Classification

Karlsruhe Institute of Technology

Methods

- Random Forest (RF)
- Random Forest (permutation)
- Gaussian Processes (GP)
- SHAP
- LIME

  Local Interpretable Model-Agnostic Explanations
- Chi-Square Test
- Maximum Relevance - Minimum Redundancy (MRMR),
- Neighbourhood Component Analysis (NCA)
- ReliefF
- ROC AUC

Diseases

- AV block
- Right bundle branch block
- Left bundle branch block

PTB-XL dataset
2000 ECGs

ECG-Features

- PR interval
- QRS duration
- T' amplitude, lead I
- ST slope, lead I
- ST slope, lead V1

Cardiologists ranking versus algorithmic ranking

Often the ranking algorithms deliver contradicting results.
Sometimes the ranking algorithm delivers results that do not match cardiologists ranking.

18HLT07 MedalCare

Metrology of automated data analysis for cardiac arrhythmia management

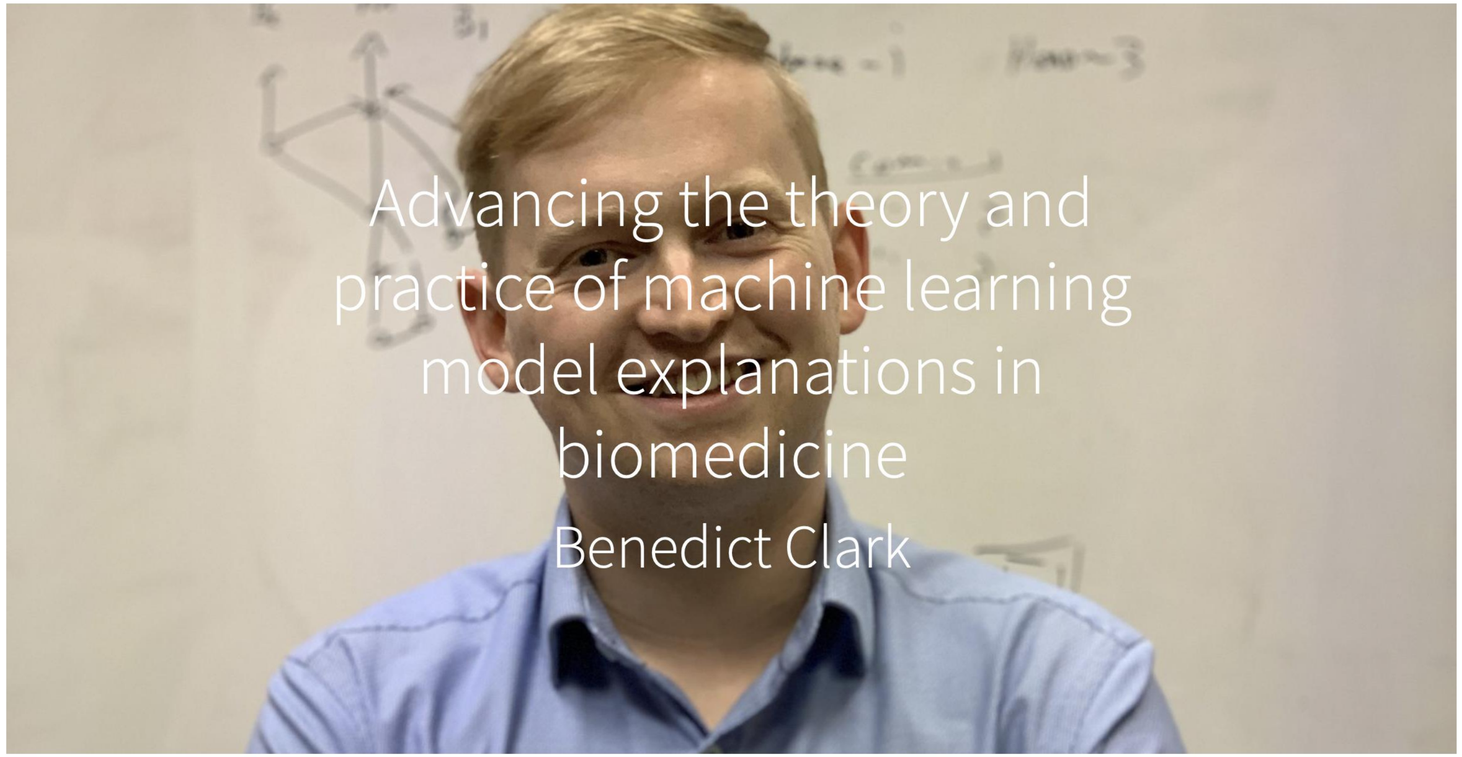
Multi-Class ECG Feature Importance Rankings: Cardiologists vs. Algorithms
Philip J Aston, Temesgen Mehari, Alen Bosnjakovic, Peter M Harris, Ashish Sundar, Steven E Williams, Olaf Dössel, Axel Loewe, Claudia Nagel, Nils Strodthoff
Computing in Cardiology 2022

M4AIM - Metrology for Artificial Intelligence in Medicine

Hans Rabus, Jörg Martin, MathMet 2022, Wednesday, November 2nd, 15:20

Welcome to M4AIM's website

Find an overview of our team and associated projects below:



PTB

https://www.m4aim.ptb.de/home/

Artificial Intelligence and Machine Learning in Medicine

- Some Examples and 4 Projects at IBT
- Some Basics of ML
- How can we Measure the Quality of an AI / ML Algorithm
- Ethical Aspects - Errors, Responsibility, Explainability and Non-Discrimination
- Regulatory Aspects
- More Data for Research - Data Protection and Reference Data
- Some Statements

Ethical Aspects of AI and ML in Medicine

- AI4People—An Ethical Framework for a Good AI Society: Opport[illegible]ties, Risks, Principles, and Recommendations, L. Floridi, J. Cowls, M. Be[illegible]i, R. Chatila, P. Chazerand, V. Dignum, Ch. Luetge, R. Madelin, U. Paga[illegible]ssi, B. Schafer, P. Valcke, E. Vayena, Minds and Machines (2018) 28:689–707 [illegible]//doi.org/10.1007/s11023-018-9482-5
- High-Level Expert Group on Artificial Inte[illegible] of the European Commission (AI-HLEG), Ethics Guidelines for Trustwo[illegible], 08.04.2019, https://ec.europa.eu/digital-single-market/en/news/ethics-g[illegible]es-trustworthy-ai
- Proposal: REGULATION OF [illegible]UROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED [illegible]ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMEND[illegible]ERTAIN UNION LEGISLATIVE ACTS (21.04.2021)

about 100 articles on Ethics of AI in Medicine

Ethics of Trustworthy AI

- **1 Human agency and oversight**
  - Including fundamental rights, human agency and human oversight
- **2 Technical robustness and safety**
  - Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
- **3 Privacy and data governance**
  - Including respect for privacy, quality and integrity of data, and access to data
- **4 Transparency - Explainability**
  - Including traceability, explainability and communication
- **5 Diversity, non-discrimination and fairness**
  - Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
- **6 Societal and environmental wellbeing**
  - Including sustainability and environmental friendliness, social impact, society and democracy
- **7 Accountability - Responsibility**
  - Including auditability, minimisation and reporting of negative impact, trade-offs and redress.

Diversity, Non-discrimination and Fairness - the Problem with „Bias“

… this is not a new topic in medicine!

- Every clinical trial must define inclusion and exclusion criteria.
- A clinical trial for all human beings is not possible.
- Results must only be applied to patients who belong to the study group.
- Sometimes important subgroups are overlooked and the patient group of the study is not perfect.

- A medical system with ML must declare an intended use (intended purpose) and show a clinical trial.
- The clinical trial must cover all patients, who are addressed in the intended purpose.
- The ML system must not be used for patients who were excluded from the clinical trial.

Responsibility and Errors that can and will happen

- Errors in implementation
  - bias
  - false classifications in the training data
  - training data set was too small
  - overfitting and leakage

The specifications promised by the producer must be met.

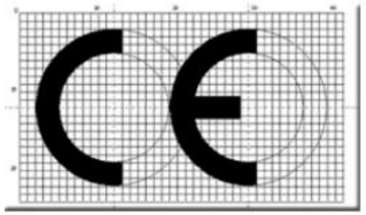
- Errors due to inherent uncertainty
  - the ground truth is not known at the moment, when the decision has to be made
  - many decisions in medicine have to be made with inherent uncertainty
  - clinical trials always show statistical uncertainty

This type of error will happen. It happens very often - even without any ML involved. ML just has to be better than any expert physician.

Artificial Intelligence and Machine Learning in Medicine

- Some Examples and 4 Projects at IBT
- Some Basics of ML
- How can we Measure the Quality of an AI / ML Algorithm
- Ethical Aspects
- Regulatory Aspects
- More Data for Research - Data Protection and Reference Data
- Some Statements

Medical Device Regulation (MDR)



- Medical devices must have the CE mark as a medical device in Europe = conformity
  - precise definition of the „intended use“
  - risk classification and risk management
  - clinical study that gives evidence about the promised properties
  - a system for vigilance and market surveillance must be operational
  - …..

Is this good enough for medical systems that contain ML?

Good Machine Learning Practice GMLP

Regulatory Aspects - the Challenges

- **We need standards adapted to medical systems containing ML (?)**
  - Definitions
  - Requirements for the quality and size of the database
  - Requirements for the separation of training, validation and test data
  - Requirements for the exclusion / avoidance of bias
  - Requirements for accuracy and uncertainty (..and number of outliers)
  - Requirements for sensitivity and robustness
  - Requirements for transparency and explainability
  - Requirements for interaction with the doctor - user interface

- …. must apply EU wide - even better: worldwide
- …. must be verifiable - quantitative measurands

FDA 10 Guiding Principles

- Multi-disciplinary expertise is leveraged throughout the total product life cycle .
- Good software engineering and security practices are implemented.
- Clinical study participants and data sets are representative of the intended patient population.
- Training data sets are independent of test sets.
- Selected reference datasets are based upon best available methods.
- Model design is tailored to the available data and reflects the intended use of the device.
- Focus is placed on the performance of the human-AI team.
- Testing demonstrates device performance during clinically relevant conditions.
- Users are provided clear, essential information.
- Deployed models are monitored for performance and re-training risks are managed.

Good Machine Learning Practice for Medical Device Development: Guiding Principles https://www.fda.gov/media/153486/download

... about the same with different priorities

EU AI Act - 21.04.2021

Establish and implement **risk management** processes (Article 9)

in light of the **intended purpose** of the AI system.

Is a medical product always a „**high-risk AI system**“?

- Use high-quality training, validation and testing data (relevant, representative, etc.) (10)
- Establish documentation and design logging features (traceability & auditability) (11 & 12)
- Ensure appropriate certain degree of transparency and provide users with information (how to use the system) (13) (explainable ?)
- Ensure human oversight (measures built into the system and/or to be implemented by users) (14)
- Ensure robustness, accuracy and cybersecurity (15)

Obligations for providers (16-28), obligations for users (29), obligations for notified bodies (30-51)

... about the same with different priorities

Artificial Intelligence and Machine Learning in Medicine

- Some Examples and 4 Projects at IBT
- Some Basics of ML
- How can we Measure the Quality of an AI / ML Algorithm
- Ethical Aspects
- Regulatory Aspects
- More Data for Research - Data Protection and Reference Data
- Some Statements

More and Quality Approved Reference Data

- Clinical reference data
  - small volume
  - very expensive
  - contain errors (ground truth?)

good example: PTB-XL database annotated 21801 clinical 12-lead ECGs from 18869 patients

- Simulated reference data
  - large volume
  - cheap
  - no problem with data protection
  - ground truth is known
  - new data sets can be created again and again
  - but they are not real patient data

10.000 simulated ECG with various diseases

18HLT07 MedalCare

Metrology of automated data analysis for cardiac arrhythmia management

Artificial Intelligence and Machine Learning in Medicine

- Some Examples and 4 Projects at IBT
- Some Basics of ML
- How can we Measure the Quality of an AI / ML Algorithm
- Ethical Aspects
- Regulatory Aspects
- More Data for Research - Data Protection and Reference Data
- Some Statements

9 Statements about ML in Medical Systems

- Machine learning in medicine will not replace the doctor, but support him or her.
- In a few exceptions, for example in emergencies, immediate action is required by the machine learning system without a doctor in the loop.
- Machine learning systems in medicine should - wherever possible - be able to explain why they came to a certain statement and indicate an „uncertainty".
- The principles of medical and product liability must be examined more closely in their application to machine learning in medicine. We need legal reliability.
- Research funding strategy in the field of machine learning in medicine is rated as “good”. Nevertheless, further efforts are necessary, for example in the areas of accuracy, transparency, explainability, quantitative evaluation.

9 Statements about ML in Medical Systems

- The protection of personal data must be guaranteed - as prescribed by the General Data Protection Regulation (EU GDPR), the Federal Data Protection Act (BDSG) and the Patient Data Protection Act (PDSG).

obvious

- Solutions are needed to generate large medical databases for research and development. The establishment of various data centers for research is welcomed. Medical data donation must be viewed as “something good”.

- It has to be clarified how clinical studies have to look like to prove the effectiveness of an ML system.

- The regulatory aspects of the approval and certification of medical devices may need to be adapted to new aspects of ML (Good Machine Learning Practice GMLP).