University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR

Why do we need Explainable Machine Learning Models?

Dr. Fani Deligianni,

fani.deligianni@glasgow.ac.uk

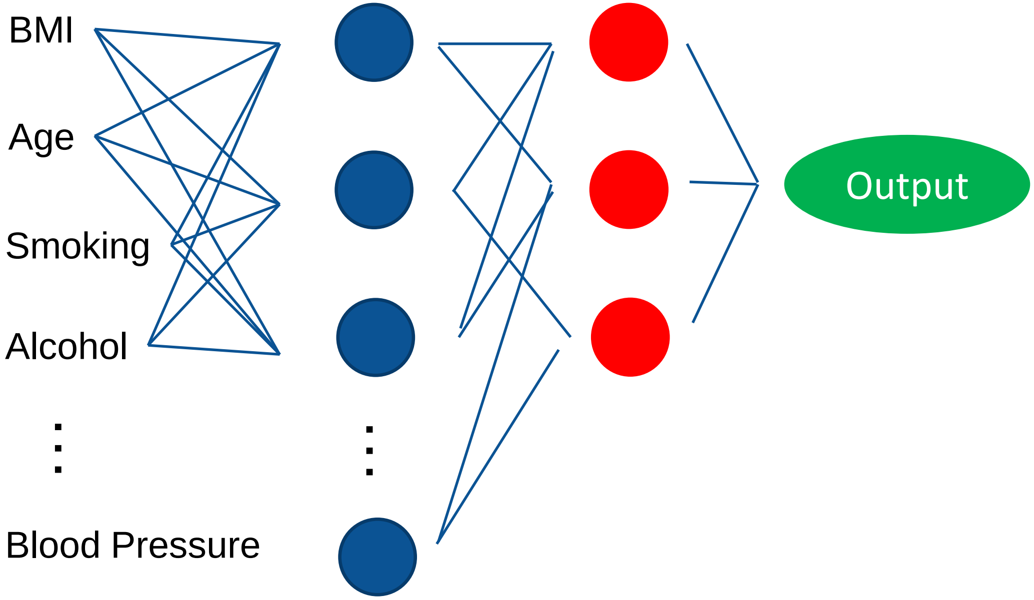
Lecturer (Assistant Professor)

Lead of the Computing Technologies for Healthcare Theme

https://www.gla.ac.uk/schools/computing/staff/fanideligianni

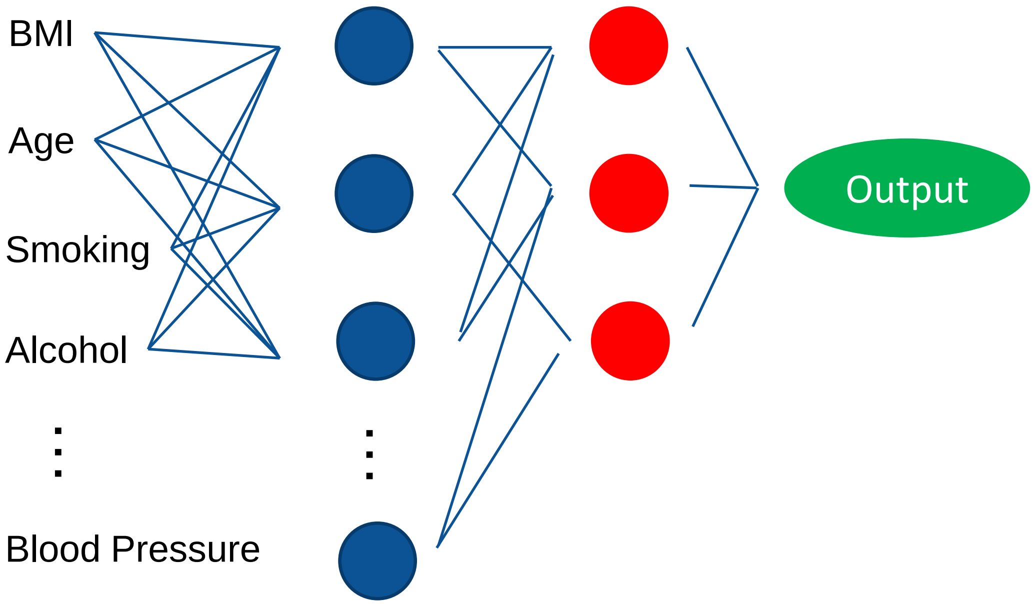
WORLD CHANGING GLASGOW

Explainable Model



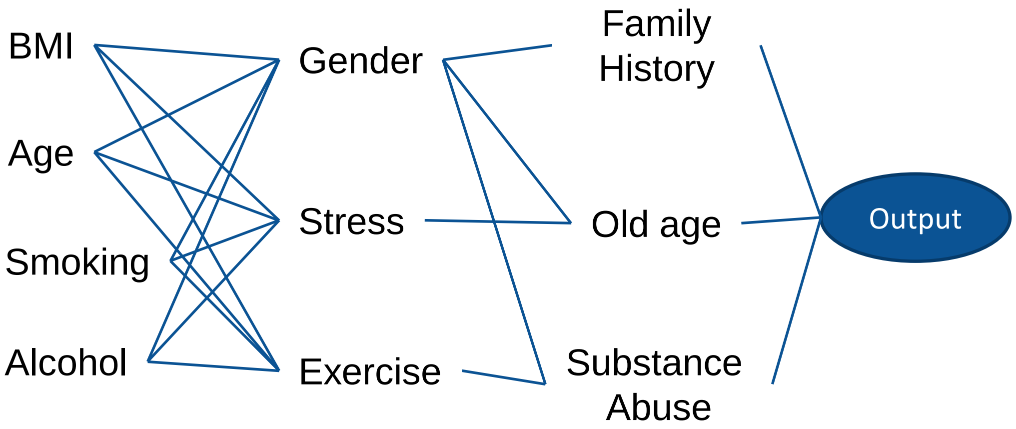
- Do we understand why the model came to this output?
- Do we know the conditions/cases that the model is successful and when it is not?
- Do we know the factors behind this output?

Explainable Model - Factors



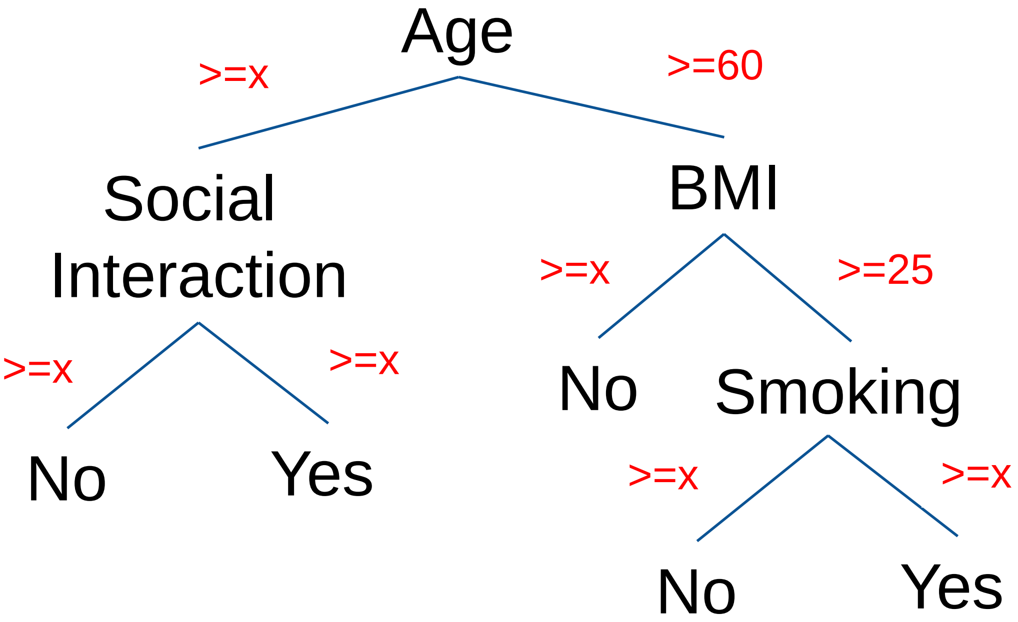
- Age is the most important factor in predicting heart failure.
- Large BMI also increases the probability of a heart attach episode
- History of smoking also increase the probability
- High blood pressure is also associated with heart failure

Explainable Model – Representation Learning



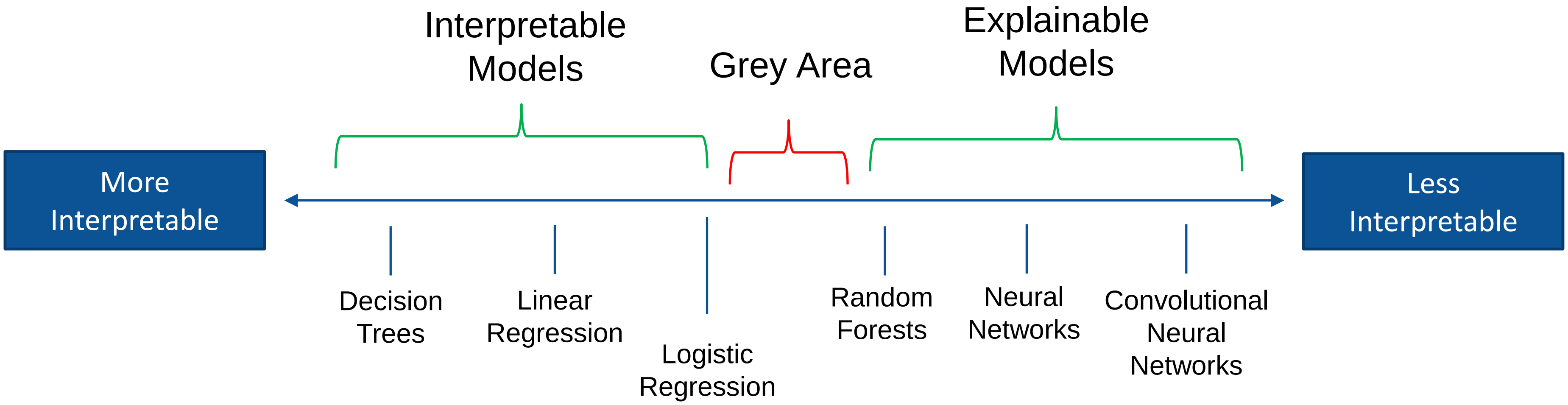
- Knowledge of the what each node represents
- Latent factors that affect the decision process
- How important each node is to the model’s performance

Interpretable Models – Decision Trees



- It is clearly what each node represents
- Easy to visualize and overview the whole decision operation
- Easy to explain to non-specialists
- Results can be tracked and associated with the output of each node

Interpretable vs Explainable Models



Interpretable vs Explainable Models

Interpretable/Transparent Models

- Model is readily understandable
- Direct Explanation
- The ability to determine cause and effect

Explainable Models

- The knowledge of which input factors are affecting the output
- The knowledge of how much they affect the decision

Interpretable vs Explainable Models

Interpretable Models

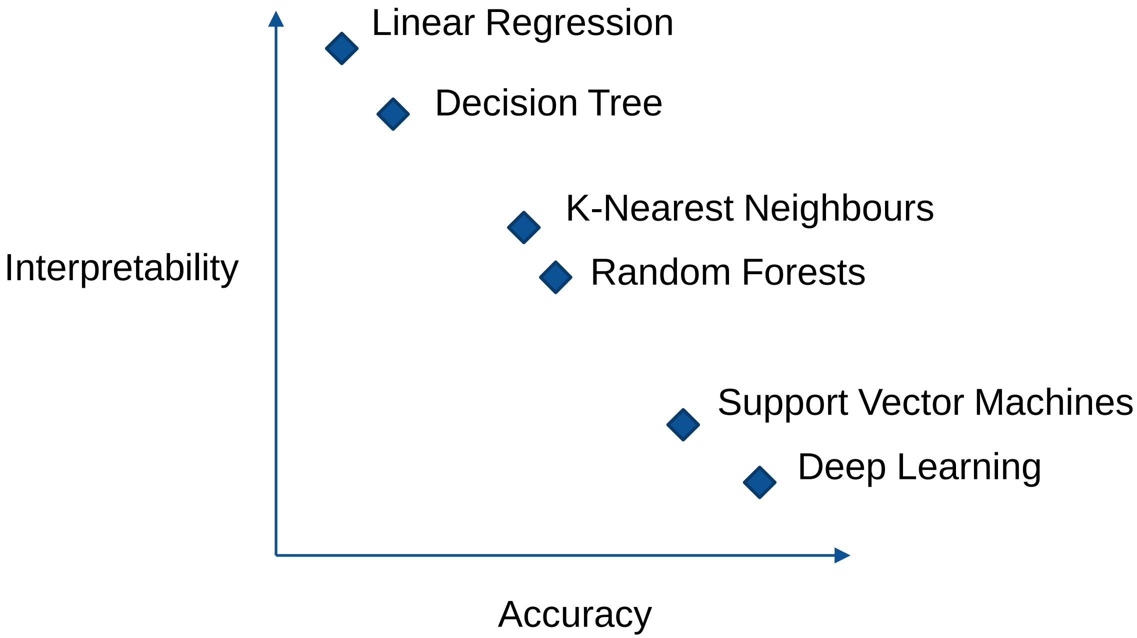
- Model is readily understandable
- Direct Explanation
- The ability to determine cause and effect

Explainable Models

- The knowledge of which input factors are affecting the output
- The knowledge of how much they affect the decision

- The ability to know what each node represents
- The ability to determine cause and effect

Interpretability vs Accuracy



Summary

- Linear models and decision trees are inherently interpretable,
- Complex models can offer better accuracy but they are inherently less interpretable
- Black boxes can be ‘explained’ in a number of different levels:
  - Based on post-hoc models that approximate their function
  - Based on local and global interpretability processes that identify which input factors are most significant and to what degree
  - Based on representation learning that identifies interpretable latent factors
- The ability to determine cause and effect

References

- Arrieta et al. ‘Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, Information Fusion, 2020.
- Molnar ‘Interpretable Machine Learning - A Guide for Making Black Box Models Explainable’ https://christophm.github.io/interpretable-ml-book/

University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR

Feature Ranking as Model Agnostic Explanations: Permutation Feature Importance

Dr. Fani Deligianni,

fani.deligianni@glasgow.ac.uk

Lecturer (Assistant Professor)

Lead of the Computing Technologies for Healthcare Theme

https://www.gla.ac.uk/schools/computing/staff/fanideligianni

WORLD CHANGING GLASGOW

Taxonomy

- Local vs **Global Explanations**
- Model Agnostic vs Model Specific Explanations
- Data Modality Specific vs Data Modality Agnostic
- Ad-Hoc vs Post-Hoc Explanations

Global Explanations

- Overall view of the model, along with data predictions and explanations.
- The **data exploration**, which displays an overview of the data set along with the prediction values.
- The **global importance**, these aggregates, features, importance values of individual data points, to show the model's overall top key.
- The explanation demonstrates how a feature affects the change in the model prediction values

Model Agnostic Approaches - Advantages

- Model Flexibility
- Explanation Flexibility
- Representation Flexibility

Model Agnostic Approaches

- **Permutation Feature Importance**
- Local Interpretable Model-agnostic Explanations
- Shapley Additive Explanations

Permutation Feature Importance (PFI)

- **Permutation feature importance (PFI)** is a model inspection technique that can be used for any fitted estimator.
- This is especially useful for **non-linear or black-box estimators**.
- The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled.
- This procedure breaks the **relationship between the feature and the target**, thus the drop in the model score is indicative of how much the model depends on the feature.

Permutation Feature Importance (PFI)

- The PFI algorithm is outlined as followed:
  - Inputs: Fitted predictive model ***m*** and dataset ***D***.
  - Compute the reference score ***s*** of the model ***m*** on data ***D*** (for instance the accuracy for a classifier or the R^2 for a regressor).

  - For each feature ***j*** and for each repetition ***k*** in **1,...,K** :
    - Randomly shuffle column ***j*** of dataset ***D*** to generate a corrupted version of the data named $D_{k,j}$.
    - Compute the score $s_{k,j}$ of model m on corrupted data $D_{k,j}$.
    - Compute importance i_j for feature f_j defined as:

$$\boxed{i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j}}$$

Permutation Feature Importance (PFI)

Algorithm 1

Algorithms for PermFIT

1: Randomly divide the data into K folds.

2: **for** $k = 1$ **to** K **do**.

3: Denote the data in k^{th} fold as V_k and the rest of the data as $\overline{V}_k$.

4: Build the machine learning model with $\overline{V}_k$, denoted as $\widehat{\mu}_k(\cdot)$.

5: **for** $j = 1$ **to** p **do**

6: Calculate $\widehat{M}_{ij}^{(P,CV)}$ for subjects in $\mathcal{D}_k$.

7: **end for**

8: **end for**

9: **for** $j = 1$ **to** p **do**

10: Calculate $\widehat{M}_{j}^{(P,CV)}$ and estimate $\widehat{\mathrm{Var}}\left[\widehat{M}_{j}^{(P,CV)}\right]$.

11: **end for**

$$\boxed{\widehat{M}_{ij}^{(P,CV)} = \sum_{k=1}^{K} \mathrm{I}(i \in V_k)\left[\left\{Y_i - \widehat{\mu}_T\left(X_{i\cdot}^{(j)}\right)\right\}^2 - \left\{Y_i - \widehat{\mu}_k(X_{i\cdot})\right\}^2\right]}$$

Mi et al. ‘Permutation-based identification of important biomarkers for complex diseases via machine learning models’, Nature Communications 2021

PFI - Disadvantages

- An in-depth understanding of the model decision is not possible
- The interaction between features via the original model is not taken into consideration
- Exact/local explanations may be required due to legal or ethical reasons

Summary

- Conceptually simple, yet powerful global ‘explainability’ method.
- PFI explains the complete dataset and not individual samples.
- It can provide a score of how important an input variable is to the prediction
- It depends on reshuffling features, adding randomness to the data measurements.

References

- Ribeiro et al. ‘Model-Agnostic Interpretability of Machine Learning’, ICML Workshop on Human Interpretability in Machine Learning, 2016.
- Mi et al. ‘Permutation-based identification of important biomarkers for complex diseases via machine learning models’, Nature Communications, 2021.

University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR

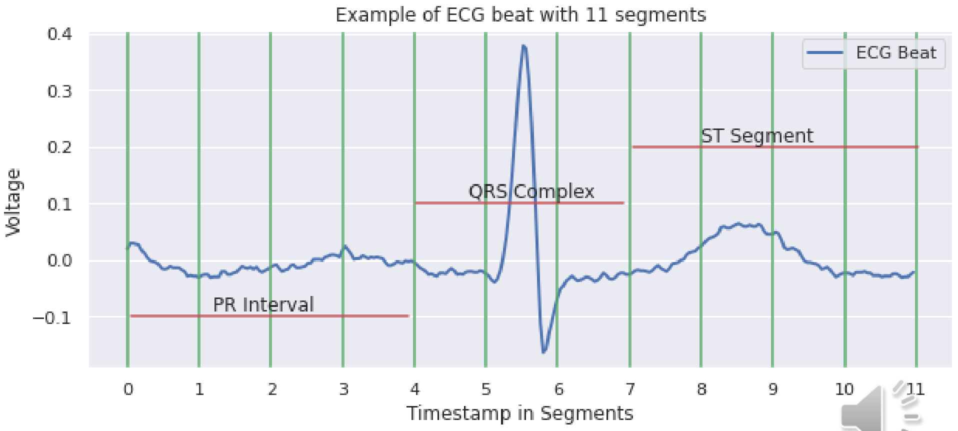
Preprocessing of ECG Signal

Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

WORLD CHANGING GLASGOW

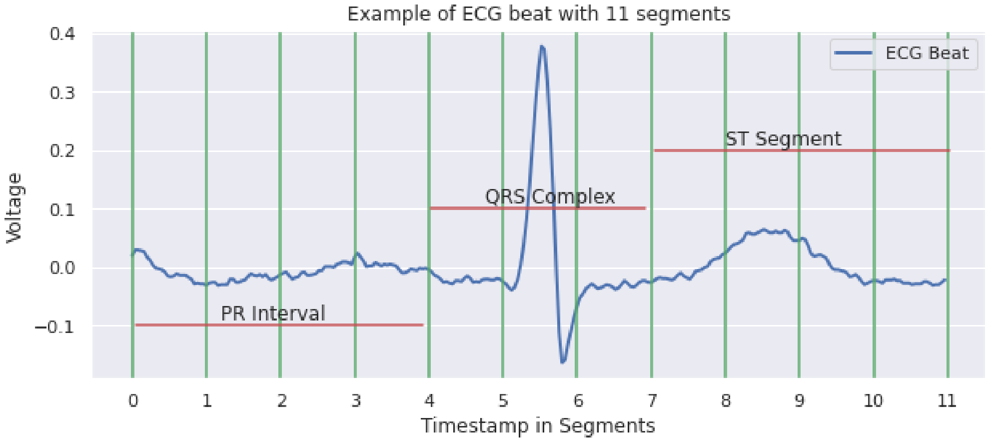
Electrocardiogram (ECG)

- An ECG test consist of collecting data through the electrical activity of the human cardiovascular system
- ECG consist of three key features which represent distinct stages of the heartbeat.
  - **P-wave:** Depolarization of the atria.
  - **QRS complex:** Depolarization of the ventricles.
  - **T-wave:** Re-polarization of the ventricles.



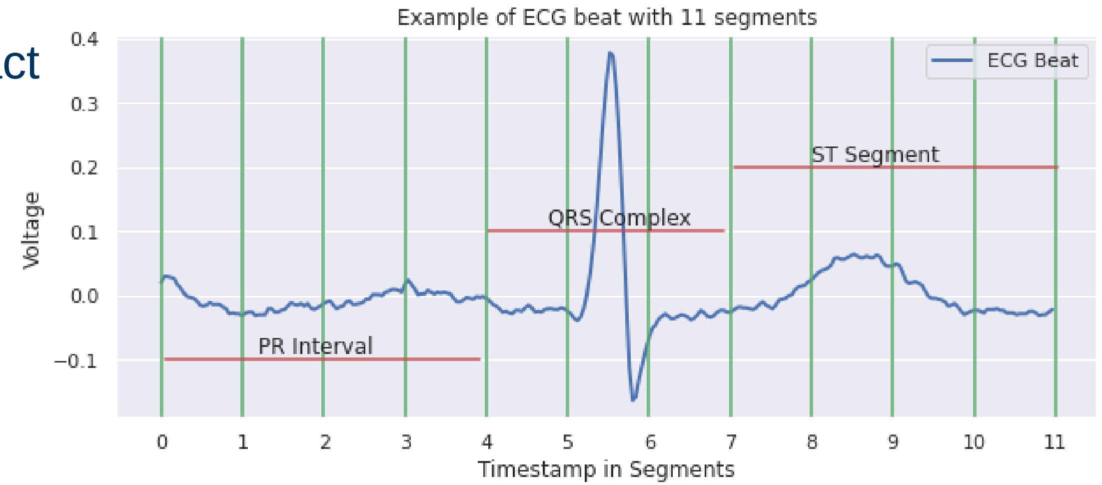
ECG Classification

- Manual ECG analysis is time-consuming and error prone
- ECG abnormalities may require continuous monitoring
- Machine learning has been extensively applied in ECG classification



Noise Interference

- The ECG signals are extremely susceptible to high and low frequency noise. These noise usually occur from:
    - Baseline wander
    - Misplaced electrode contact
    - Motion artifacts
    - Power line interference



MIT-BIH ECG Dataset

- The MIT-BIH dataset used for this investigation is a public database consisting of a large number of annotated beats.
- It is frequently used for time-series classification research.

- The MIT-BIH Arrhythmia Database contains sections of ambulatory ECG recordings:
  - From 47 subjects, digitized at 360 samples per second per channel.
  - 11-bit resolution at 10-mV range on two channels.
  - Here 23 recordings were picked at random from a set of 4000 24-hour ECG recordings.
  - Collected from a population 60% of inpatients and 40% outpatients.

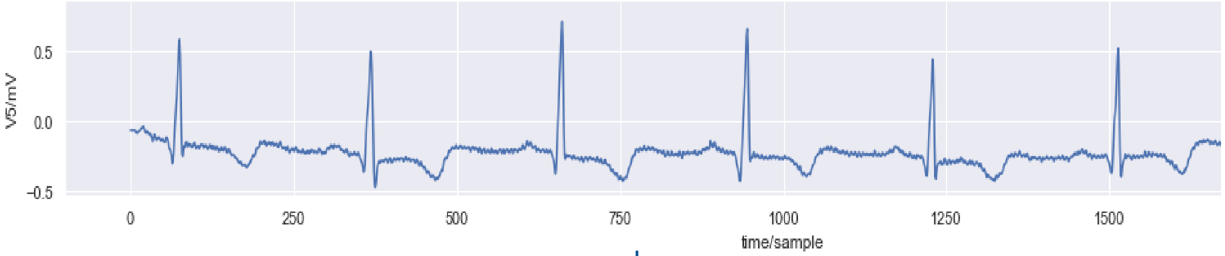
MIT-BIH ECG Dataset

- This data has been pre-annotated and labelled by cardiologists.
- These different annotations refer to various normal and abnormal ECG signals which represent different types of arrhythmia.
- The dataset consists of ECG signals of various classes, but the eight classes used for this investigation are ’N’, ’L’, ’R’, ’V’, ’A’, ’F’, ’f’, ’/’.
- The table shows the description and numerical identification values assigned to these classes.

Class	ID	Beat Description
N	1	Normal
L	2	Left Bundle Branch Block
R	3	Right Bundle Branch Block
V	4	Premature Ventricular Contraction
A	5	Atrial Premature
F	6	Fusion of Ventricular and Normal
f	7	Fusion of Paced and Normal
/	8	Paced

Data Pre-processing

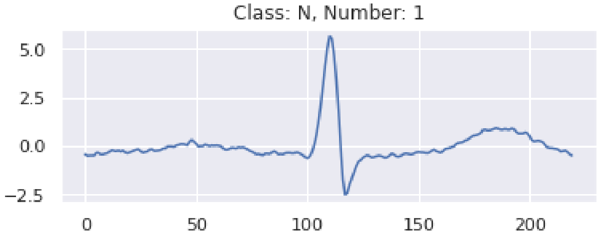
Raw Data

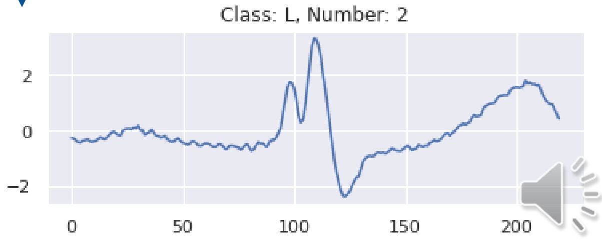


Data Pre-processing

Individual Beat, Centre R-peak, Standardize, Beat Annotation

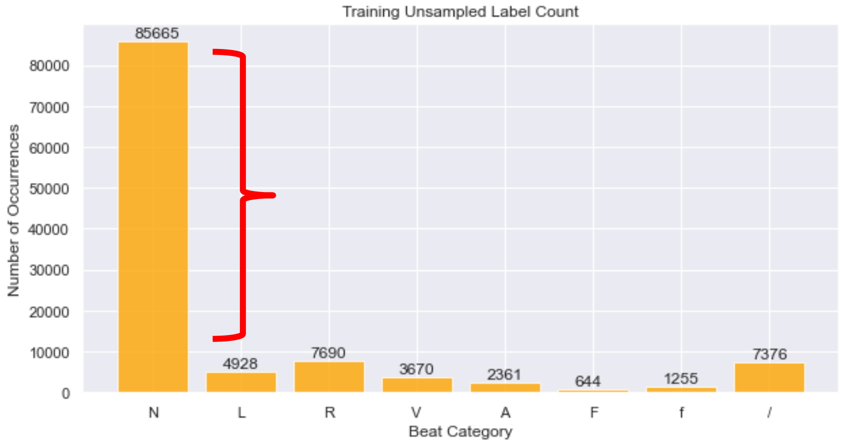
Input Data

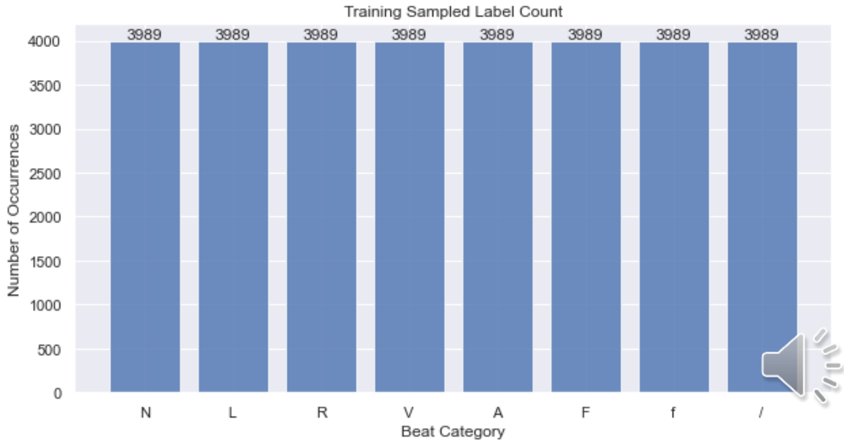




Class Imbalance

- The normal class is over-represented in the data
- Resampling is based on a **bootstrap method** which resamples a dataset with replacement, iteratively
- For **up-sampling and down-sampling**, the sample value was calculated by taking the mean values of the total number of beats of the abnormal classes.





Summary

- Preprocessing of the ECG signal include:
    - Filtering to remove noise
    - Annotation of the R-peaks
    - Segmentation of the recordings into ECG beats
    - Resampling the data to address the imbalance problem

References

- Mark RG et al. ‘An annotated ECG database for evaluating arrhythmia detectors’, IEEE Transactions on Biomedical Engineering 29(8):600, 1982
- Moody et al. ‘The impact of the MIT-BIH arrhythmia database’, IEEE Engineering in Medicine and Biology Magazine 20(3), 45-50, 2001
- Yola et al. ‘Improving ECG Classification Interpretability using Saliency Maps’, IEEE BIBE, 2020.

University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR

Explainability Use-Case

Dr. Fani Deligianni,

fani.deligianni@glasgow.ac.uk

Lecturer (Assistant Professor)

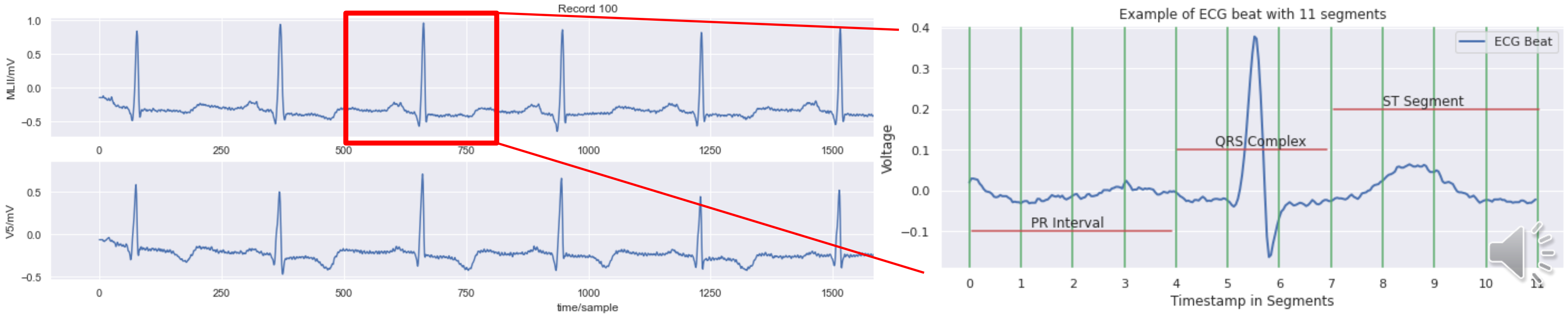
Lead of the Computing Technologies for Healthcare Theme

https://www.gla.ac.uk/schools/computing/staff/fanideligianni

WORLD CHANGING GLASGOW

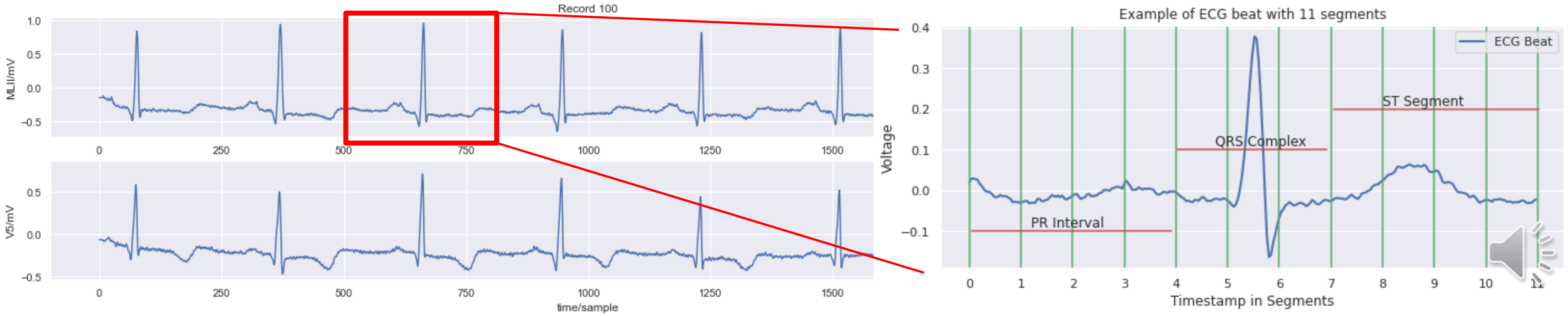
Application of PFI in ECG Classification

- The ECG beats were divided into slices of 11 segments.
- This helped interpret which segment is being given more importance by the classifier.
- The slices were made by replacing the data points with the average point for each slice.

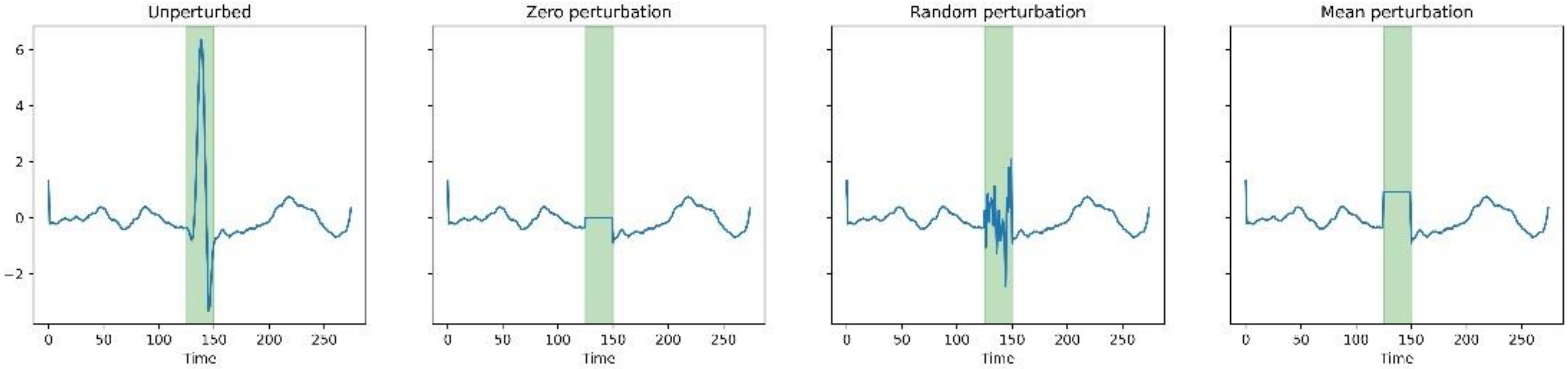


ECG Segmentation

- Segments 1-4 cover the PR interval.
- Segments 5-7 cover the QRS complex
- Segments 8-11 cover the ST segment.
- We expected to see the model focusing on important morphological features of the ECG beat, such as the PR interval, the QRS complex, and the ST segment.



PFI for ECG Classification



Assessment Tasks

- Inspect your data and plot different types of arrythmia. Run the python notebooks provided and plot also the distribution of samples across classes

4-6 members per group: (At least **two** different classifiers)

- Classification of ECG beats based on the holdout splitting method
- Classification of ECG beats based on the leave-out, patients-hold out validation protocol
- For each of the models developed above use permutation feature importance to explain the model's function
- Apply the same explainability technique with different type of classifiers and discuss the differences

6 members per group: (In addition to the above task):

- Use at least one clustering technique to visualize the data and understand better their structure and how well classes are separated