Highlights

Interpretable Heartbeat Classification using Local Model-Agnostic Explanations on ECGs

Inês Neves,Duarte Folgado,Sara Santos,Marília Barandas,Andrea Campagner,Luca Ronzio,Federico Cabitza,Hugo Gamboa

- We present an in-depth study on the technical feasibility and practical usefulness of visual explanations for ECG classifiers
- We propose using the time series derivate to support state-of-the-art XAI methods measuring feature importance considering the temporal domain
- We conducted an informative user study to evaluate the potential of visual explanations on ECGs

Interpretable Heartbeat Classification using Local Model-Agnostic Explanations on ECGs

Inês Neves[a,1], Duarte Folgado[a,b,1], Sara Santos[a], Marília Barandas[a,b], Andrea Campagner[c], Luca Ronzio[c], Federico Cabitza[c] and Hugo Gamboa[a,b]

[a] *Associação Fraunhofer Portugal Research, Rua Alfredo Allen 455/461, 4200-135 Porto, Portugal*

[b] *Laboratório de Instrumentação, Engenharia Biomédica e Física da Radiação (LIBPhys-UNL), Departamento de Física, Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal*

[c] *Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Viale Sarca 336 – 20126 Milano, Italy*

ARTICLE INFO

Keywords:
machine learning
time series
heartbeat classification
electrocardiogram
explainable artificial intelligence
model-agnostic method
visual explanations
usability
human–AI interfaces

ABSTRACT

Treatment and prevention of cardiovascular diseases often rely on Electrocardiogram (ECG) interpretation. Dependent on the physician's variability, ECG interpretation is subjective and prone to errors. Machine learning models are often developed and used to support doctors; however, their lack of interpretability stands as one of the main drawbacks of their widespread operation. This paper focuses on an Explainable Artificial Intelligence (XAI) solution to make heartbeat classification more explainable using several state-of-the-art model-agnostic methods. We introduce a high-level conceptual framework for explainable time series and propose an original method that adds temporal dependency between time samples using the time series' derivative. The results were validated in the MIT-BIH arrhythmia dataset: we performed a performance's analysis to evaluate whether the explanations fit the model's behaviour; and employed the 1-D Jaccard's index to compare the subsequences extracted from an interpretable model and the XAI methods used. Our results show that the use of the raw signal and its derivative includes temporal dependency between samples to promote classification explanation. A small but informative user study concludes this study to evaluate the potential of the visual explanations produced by our original method for being adopted in real-world clinical settings, either as diagnostic aids or training resource.

1. Introduction

According to the World Health Organisation [1], cardiovascular diseases are responsible for 31% of worldwide deaths each year. Since cardiovascular diseases are the leading cause of global deaths, their treatment and prevention rely on monitoring data and pattern evolution on patients to develop cost-effective health care innovations. Electrocardiogram (ECG) monitoring is a regular exam for the triage and diagnosis of cardiovascular conditions as it is an inexpensive and non-invasive procedure by which to assesses the heart function through its electric activity. In these times, the cardiologist or the cardiac technician manually analyse the ECG, a task that is prone to subjective errors and observer's variability [2].

Although the implementation of Machine Learning (ML) models that assist heartbeat classification is thriving in the healthcare sector, they are still in the early stage of their adoption [3]. In 2017, a survey conducted in 85 hospitals showed that approximately 5% had the intention to adopt Artificial Intelligence (AI) solutions to this aim, most of which said to be highly uncertain about when to start deploying them. The barriers to adopt these models are mainly due to non-existing executive and physician technology buy-in and lack of trust [4, 5].

AI systems can support ECG analysis in virtue of their capability to process large amounts of data, detect patterns, and make accurate predictions; however, their general lack of interpretability can back up resistance attitudes to adoption [6] and hamper the acceptance of these models in medical diagnosis, where accountability (and liability) are crucial [7].

*Corresponding author

duarte.folgado@fraunhofer.pt (D. Folgado)

ORCID(s): 0000-0002-8481-6079 (D. Folgado); 0000-0002-2554-3648 (S. Santos); 0000-0002-9445-4809 (M. Barandas); 0000-0002-4065-3415 (F. Cabitza); 0000-0002-4022-7424 (H. Gamboa)

[1] These authors contributed equally.

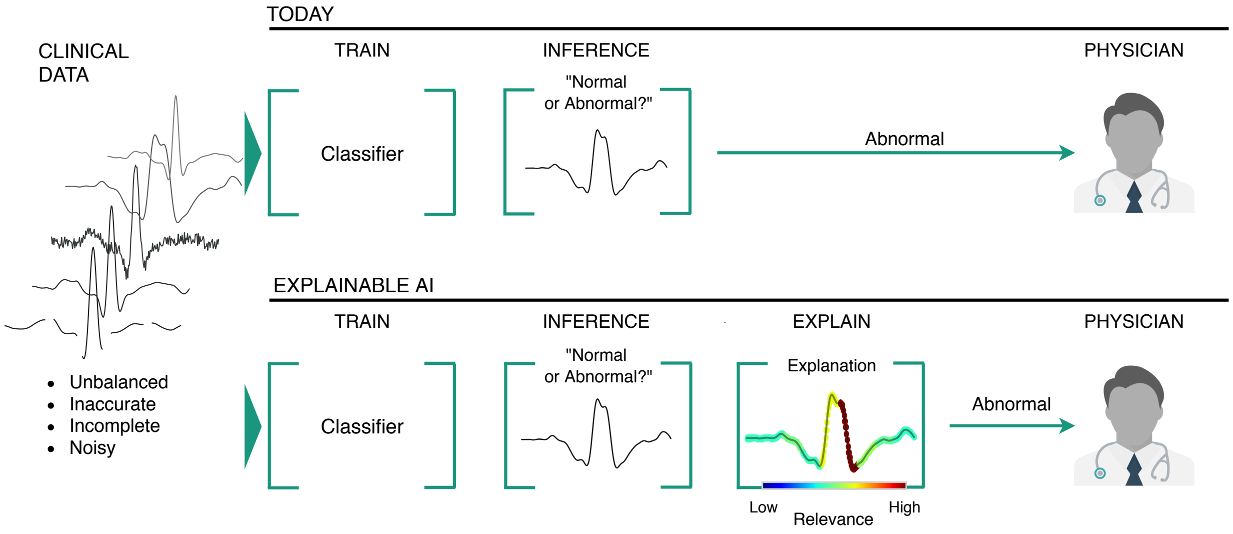


Figure 1: Comparison between the machine learning traditional pipeline (on top) and the XAI pipeline (bottom), that introduces explanations through more "explainable" models.

However, making ML classifiers more interpretable is far from being an easy task. For example, to "explain" the predictions yielded by decision trees, one might consider following the tree path. Nonetheless, a high number of nodes could make this process cumbersome and related explanations almost incomprehensible to humans [8]. In addition, integrating ML models in clinical support systems is a challenging task by itself, since these models have to ground on uncertain, imbalanced, heterogeneous, and noisy datasets [9, 10], which present a high number of features and that yet are often not large enough to allow the precise modelling of the complex systems of the patients. From the regulatory point of view, the European Union regulations concerning the treatment of personal data have recently opened the debate over the "right to explanation" of patients and other stakeholders, when a decision is based (or coincides) with an automatic process, and urged the scientific community to develop ways to guarantee such a right, or at least provide healthcare practitioners with *explanatory descriptions* backing up the AI decisions [11].

In what follows, we will rely on an intuitive notion of "explanation", a term that in this context denotes any indication that could help the human decision-maker (in our case, the ECG reader) *understand* the output of the decision support (in our case an ML classifier).

Nowadays, it has become essential to endow AI systems with the capability to provide not only accurate diagnoses but also additional information that explain, or justify, the decision of complex classifiers [12], as illustrated in Figure 1. However, Explainable Artificial Intelligence (XAI) studies are still in their infancy, and so far, there is a lack of consensus on the best appropriate theoretical and algorithmic methods and standard assessment metrics [13]. Furthermore, most medical XAI research works have primarily focused on computer vision tasks and less often on time series. With this work, which focuses on this latter type of data, we aim to fill this gap in the specialist literature.

In particular, we will address the following questions: what are the most appropriate means to explain the classifications of a time series classifier? Would it be feasible to transpose XAI methods that have been validated for diagnostic images into the context of time series? What are the main requirements for the automatic production of sound explanations of the decisions of a time series classifier?

A first approach to be evaluated would employ what we call *visual explanations*, that is, visual attention mechanisms that superimpose chromatic clues on each time series, inspired by salience maps in 2D images, in order to highlight correlations between samples and between features, or show the relevance of the instance's morphology [14].

Although standard XAI models can, at least in principle, be adapted to temporal data, these generally assume independence among the features, which implies temporal independence. The dependence between samples is an inherent characteristic of time series, and this latter should not be omitted in the visual explanation. Therefore, we argue that in the context of time series classifiers, designers must consider the temporal dependency between samples or multiple time series.

This work presents an in-depth study on the technical feasibility and practical usefulness of visual explanations for ECG classifiers. We also propose a concise, high-level taxonomy towards the standardisation of XAI methods for time series, and an adaptation of state-of-the-art methods optimised for time series data. We applied our approach to a well-known public ECG database - the MIT-BIH arrhythmia dataset. Finally, we describe a short user study to validate our approach and evaluate the practical usefulness of the proposed visual explanations in a simulated yet realistic setting, involving three ECG readers of different backgrounds and expertise.

Section 1 highlights our motivation and outlines the context of our study. Section 2 reports previous work done on XAI applied to clinical time series classification. Section 3 presents our framework for temporal data classification, by illustrating different XAI methods. Sections 4 and 5 report and discuss the findings, respectively. Section 6 presents our small empirical user study. Finally, Section 7 expresses our main final remarks and conclusions.

2. Related Work

The importance of explaining machine decisions has been present since the early days of AI. The first generation of expert systems, which emerged in the 1970s, relied on a set of inference rules, a network of relationships, and, in some circumstances, the introduction of probabilities. One of the most notable expert systems on the medical domain was entitled MYCIN [15]. The system contained an explanation mode, which allowed MYCIN to explain its conclusions and the reasoning process of its inference engine.

Those earlier models were shallow in comparison with state-of-the-art approaches and potentially easier to explain. From 2010 onwards, the implementation of more complex models with the dissemination of deep learning methods has reinforced the need for model explanation techniques [16].

Most of the current research in XAI has focused on developing methods for computer vision and natural language processing tasks. Therefore, there are relatively few works on literature that systematically address time series data and even less prevalent in clinical time series. Table 1 summarises the most recent works which address XAI in the context of clinical time series.

One of the criteria used to classify XAI methods is related to model specificity. Model-specific methods are limited to specific classifiers as their development relies on such model's internals. Model-agnostic approaches can be used on any model and are applied after the model has been trained. They consider the model as a black-box and explain its behaviour usually by analysing pairs of input and output. Model-specific methods were used in the clinical domain to explain deep learning approaches applied to time series in case-based reasoning technique using prototypes [17], attention mechanisms [18, 19] and the Layer-wise Relevance Propagation (LRP) [20].

A model-agnostic approach was explored by Mujkanovic [21] and Guillemé et al. [22]. Both works introduced adaptations of state-of-the-art XAI methods, such as SHappley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) into explaining time series. LIME and SHAP are both perturbation-based approaches that provide explanations based on the variation of the output after applying perturbations on the sample's vicinity. A disadvantage of perturbation-based methods is that they omit temporal dependencies, producing explanations only limited verifiable on time series data. These methods produce local explanations for single predictions through which the produced variation represents a relevance score for each sample. They mostly differ in the used approach to calculate the relevance score. SHAP is based on game theory to calculate Shapley Values as relevance's scores [23] and LIME uses a surrogate linear model to explain local perturbations [24].

To assess the quality of the explanations, Doshi-Velez et al. [13] proposed an evaluation taxonomy, starting from application-grounded evaluation, the most specific and costly, requiring humans with expert knowledge; to the least specific, functionally-grounded evaluation, requiring a formal definition of interpretability as a proxy for explanation quality and exempting human tasks.

There is limited work on validation strategies to characterise XAI methods for time series, without human intervention. Once again, Guillemé et al. [22] proposed a method to measure fidelity by comparing an explanation from the proposed XAI method against an interpretable classifier, using shapelets. Shapelets are the most discriminative subsequences from a time series, to identify classes. These shapelets can be used in a tree-based classifier to make predictions based on the distance between the shapelet and subsequences of the instance [26]. On the other hand, Mujkanovic [21] compared the explanations from the adapted SHAP from different time series classifiers, through the median correlation. Furthermore, Schlegel et al. [25] presented a study on metrics to assess explanations of time series classifiers, by analysing the decrease in the model's performance, after the most relevant sequences calculated by the XAI method are replaced. The greater the drop in model's performance, the more reliable and representative

Table 1
Description of recent XAI studies on medical temporal data.

Study	Objective	Method	Specificity	Data
Song et al. (2017) [18]	Explain medical time series classification from a deep learning classifier	Attention	Model-specific	MIMIC III
Lin et al. (2019) [19]	Explain a myotonic dystrophy prediction of a deep learning classifier	Attention	Model-specific	Electromyogram from handgrip movement
Gee et al. (2019) [17]	Explain deep learning classifiers using a set of time series data	Learned prototypes	Model-specific	Neonatal ICU
Horst et al. (2019) [20]	Explain the deep learning prediction of individuals from gait analysis	LRP	Model-specific	Gait data
Guillemé et al. (2019) [22]	Explain time series classification	LIME SHAP	Model-agnostic	UCR
Mujkanovic et al. (2019) [21]	Compare the explanations of different time series classifiers	SHAP	Model-agnostic	UCR
Schlegel et al. (2019) [25]	Assess XAI methods for time series	LIME SHAP LRP DeepLIFT	Model-agnostic Model-specific	MIT-BIH arrythmia UCR

the explanation is. The used substitutions were various, such as replacing by zero, the mean value, or swap, which requires swapping the order of the samples amongst that sequence. The last replacement allows us to evaluate if temporal dependency is being taken into account by the XAI method. This research showed that LIME provides the least reliable results when compared to SHAP and other model-specific methods.

The related work highlights the latest effort in including XAI methods into clinical time series classifiers and exposes the lack of model-agnostic applications. Furthermore, the existing model-agnostic methods are still far from optimal as they assume temporal independence. To tackle the challenge of sample independence in this type of explainable method, the instance's derivative is introduced to force the XAI method to gain temporal sensitivity. Moreover, it becomes crucial to define a system for categorising time series explanations and a validation protocol to enable our study's support. Accordingly, a rigorous evaluation of the chosen dataset is relevant to creating explanation methods able to provide explanations consistent with the complex model and produce meaningful justifications for domain experts.

3. Explainable Artificial Intelligence on Time Series

In this section, we summarise some concepts related to the application of XAI in the time series domain. Firstly, we start by introducing some formalisms. Then we propose a short explanation taxonomy for time series methods and lastly, we present some methods and respective adaptation to explain time series. In Table 2 we briefly list the relevant terminology, concerning ECGs and time series, for the present paper.

Definition 1. A time series or instance $X = \{x_1,\ x_2, ...,\ x_n\}$ is an ordered set of samples with equally spaced sampling, where n represents the number of samples.

Definition 2. A dataset $D = \{X_1, X_2, ..., X_N\}$ is a collection of N time series.

Quite often we are interested in evaluating the performance of the machine learning model on data that it has not seen before, since it determines how well the model is generalising to unseen data. We evaluate these performance

Table 2
Basic dictionary of terms relevant to ECG and time series, as used in the present paper.

Term	Definition
Electrocardiogram (ECG)	A consecutive sequence of heart-beats from a given patient
Instance (syn. Time Series)	A segment of a ECG associated with a single heart-beat An ordered set of equally spaced samples
Sample	A single data-point in an instance
Window	A subsequence of an instance

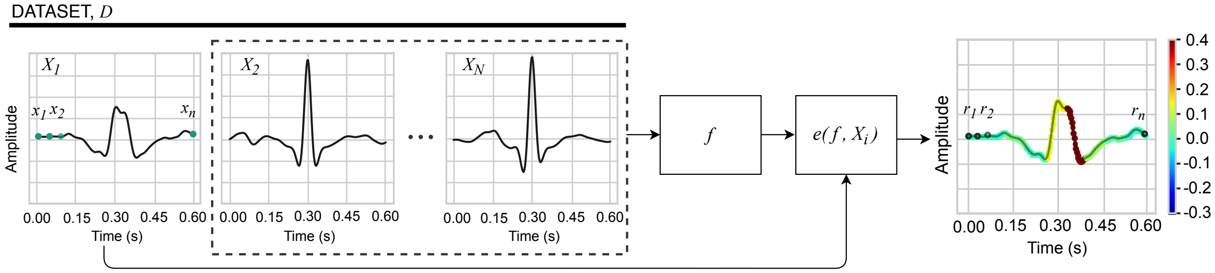


Figure 2: Incorporating explanation in time series. A prediction is explained by a function e that measures the relevance of each sample for the classifier.

measures using a separate portion of data that was not used during training that we denote as the test set. Therefore, the dataset D is composed of two distinct sets of data: training, D_{train}, and test set, D_{test}, such as $D = \{D_{train} \cup D_{test}\}$.

Definition 3. A prediction $\hat{y}$ is obtained through $\hat{y} = f(X)$, where $f(X)$ represents the classifier.

For model-agnostic methods, a local explanation that explains a single prediction by measuring relevance can be formally defined as:

$$R_i = e(f, X_i), \quad i \in \{1, ..., N\} \tag{1}$$

An instance X_i is explained by an explanation method e, which depends on both the classifier and the instance being explained. The function e yields scores for each sample corresponding to their relevance on the prediction. Figure 2 summarises the standard model-agnostic explanation pipeline: the i^{th} instance from D_{test} is predicted according to the classifier $f(X_i)$. Afterwards, an explanation method e attributes relevance scores $R = \{r_1, r_2, ..., r_n\}$ for each sample of i.

3.1. Taxonomy

Although there are some works in the literature dedicated to the taxonomy of XAI [27, 28, 29, 30], none of them focuses particularly on the domain of time series. A clear and objective taxonomy is essential to stimulate research directions and promote standardisation in the community for future work.

We propose dividing the XAI methods according to three levels of time series interpretation:

- **Sample-based methods:** the classifier's predictions are explained by attributing relevance values to the raw time series samples;
- **Feature-based methods:** the classifier's predictions are explained by attributing relevance values to the features extracted from the time series. Time series are often divided into windows before the feature extraction process, therefore, explanations refer to the relevance of a given window to the classifier's outcome;

- **Morphology-based methods:** the classifier's prediction is explained by the relevance of morphology attributes. The attributes are often perceived by data visualisation and comprise rising and falling slopes, concavity, direction, frequency, time, and amplitude range of a slope, among others [31].

- **Text-based methods:** the classifier's decision is explained by means of textual descriptions generated on the basis of the content of the original written reports that are associated with the exams of the training set and ground truth [32].

At the first level, which will be the main focus of this paper, sample-based explanations are issued in raw time series. The classifier's prediction is explained by the impact that each sample or window has on a particular decision. This relevance is expressed as a numerical weight. Positive weighted samples contribute towards the classification of the complex model, and negative weighted samples contribute contrarily to the model's prediction. This approach is viable in the binary and the multiclass classification as it allows us to simplify the problem into whether the sample has contributed or not to the final prediction.

Due to the high computational cost of most explanation methods, it is often convenient to explain a group of samples.

Definition 4. A window W is a subsequence of X, such as $X = \{W_1, W_2, ..., W_t\}$. X can be represented as a set of windows such as $W_1 = \{x_1, ..., x_l\}$, $W_2 = \{x_{l+1}, ..., x_{2l}\}$, $W_t = \{x_{(t-1)(l+1)}, ..., x_{tl}\}$ and l is an arbitrary window size.

3.2. Methods

In this section, we present three model-agnostic methods which could be described, in terms of the previously mentioned taxonomy of time series explanation methods, as sample-based. Several prior works addressed the problem of explaining a classifier using these methods in several applications and data types [33, 34, 35, 36, 37, 38]. For each method, we clearly describe our proposed adaptations so that they can adequately be used with time series data.

3.2.1. Permutation Sample Importance

Permutation Feature Importance (PFI) is a perturbation model-agnostic method that provides global interpretability by inspecting the model score after a single feature value is randomly shuffled [39]. The increase or drop in the model score describes the relationship between the prediction and the permuted feature. PFI replaces each feature p times for other features from randomly picked instances from the dataset. Firstly, the reference score of the classifier is computed. Each feature is shuffled, generating a perturbed version of the test dataset, D_{test}. The recalculated model score corresponds to the model score when using the permuted dataset. Finally, the importance of each feature is given by the difference between the initial model score and the average of the model score with the permuted data, previously repeated p times [40]. The higher the drop in the model's score, the more relevant is the feature [41].

We propose a similar approach to time series data. For convenience we denote the adaptation as Permutation Sample Importance (PSI), to clearly denote that is applied to the raw time series and not over extracted features from time series. The permuted dataset, $\tilde{D}$, in this case, is composed of several copies of the time series we seek to explain, with the samples permuted. The permutation is applied to each sample, and the samples are replaced for the values of randomly chosen time series from D, that belong to the opposite class or classes, considering the multiclass situation. This step forces the permutation to generate a new instance in which the permutations are from a different class we are intended to explain.

The relevance of a sample from the instance i, r, is defined as the mean of the differences between the posteriori probabilities of the baseline and permuted dataset, as defined according to Equation 2:

$$r = \frac{1}{p} \sum_{j=1}^{p} P(\hat{y}_i | D) - P(\hat{y}_i | \tilde{D}) \tag{2}$$

The baseline probability $P(\hat{y}_i | D)$ is the predicted probability of the classifier for the instance X_i. p is the number of times the permutation is applied to each window.

3.2.2. Local Interpretable Model-agnostic Explanations

LIME [24] is a perturbation-based method that uses a surrogate interpretable model to locally replace the complex model, providing local interpretability. The surrogate model is trained with the predictions given by the complex classifier, of a perturbed dataset weighted around the instance being explained. LIME ensures both interpretability and local fidelity by minimising how unfaithful is the local approximation of the surrogate model, g, to the complex classifier, f.

The explanation, R, produced by LIME is obtained according to Equation 3:

$$R = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_{X_i}) + \Omega(g) \tag{3}$$

X_i refers to the instance being explained, G denotes the different families of interpretable models, $\mathcal{L}(f, g, \pi_{X_i})$ is the fidelity function, measuring the reliability of the approximation provided by the interpretable model in the vicinity defined by π_{X_i} and, $\Omega(g)$ denotes the complexity of the interpretable model. The perturbed instances are weighted according to an exponential kernel, π_{X_i}, that attributes higher weights to instances similar to X_i, given by Equation 4:

$$\pi_{X_i} = e^{-\left(\frac{d}{\sigma}\right)^2} \tag{4}$$

Where d corresponds to a chosen distance metric and σ is the kernel's width. The kernel defines the meaningful area around the instance being explained and its width the size of the neighbourhood. It has been noted, by Kopper and Molnar in [42] and Laugel et al. in [43], that choosing an adequate value for the scale paremeter σ is of critical importance to ensure adequate approximations.

The perturbed dataset is generated in the interpretable data representation. This representation indicates the presence or absence of a given element. For instance, in text classification, denotes the absence of a word, and in image classification represents the presence or absence of a contiguous patch of similar pixels (a super-pixel). Given $x \in \mathbb{R}^d$, the perturbed sample is denoted as $z' \in \{0, 1\}^d$. Perturbations occur through the random attribution of 0 in different features, and the relevance of each is calculated by how much their removal has changed the prediction.

The interpretable model learns over the perturbed dataset. Popular choices for the interpretable model are (regularized) linear models (such as Lasso regression), decision tree or decision rules. In this paper, we consider the first approach, hence our implementation of LIME is based on (standard) Lasso regression, for which the optimization problem in Equation 3 reduces to:

$$argmin_{\beta} \sum_{z} \pi_{X_i}(z)(\hat{y}_z - \beta \cdot z)^2 + \lambda \, \|\beta\|_1 \tag{5}$$

where β is the vector of weight coefficients, λ is a regularization parameter, X_i is the instance to be explained, z is a perturbed instance and $\hat{y}_z$ refers to the probability score predicted by the classifier f to be explained for the perturbed instance z. In this case the (absolute values of) weights of the linear model, learned via a least-squares procedure, denote the relative importance of each feature. Higher coefficients imply a higher impact on the prediction and lower weights represent features that have a smaller impact on the model's decision.

LIME tries to ensure that the linear model is locally faithful, i.e., it must correctly approximate the complex classifier in the vicinity of the instance being explained. Whilst there is no standardised methodology to evaluate faithfulness, the performance metrics and the R^2 coefficient of the linear classifier are evaluated. A correctly local prediction with a high R^2 suggests that the linear model is correctly approximating the complex model for a given instance.

For time series, we propose adapting the LIME original idea. Instead of perturbing features, we perturb time series' windows. The perturbed window $\tilde{W}$ are given as follows:

$$\tilde{W}_w = \begin{cases} W_w, & \text{if } z'_w = 1 \\ p(W_w), & \text{if } z'_w = 0 \end{cases} \tag{6}$$

Where W_w is the $w - th$ window of the time series.

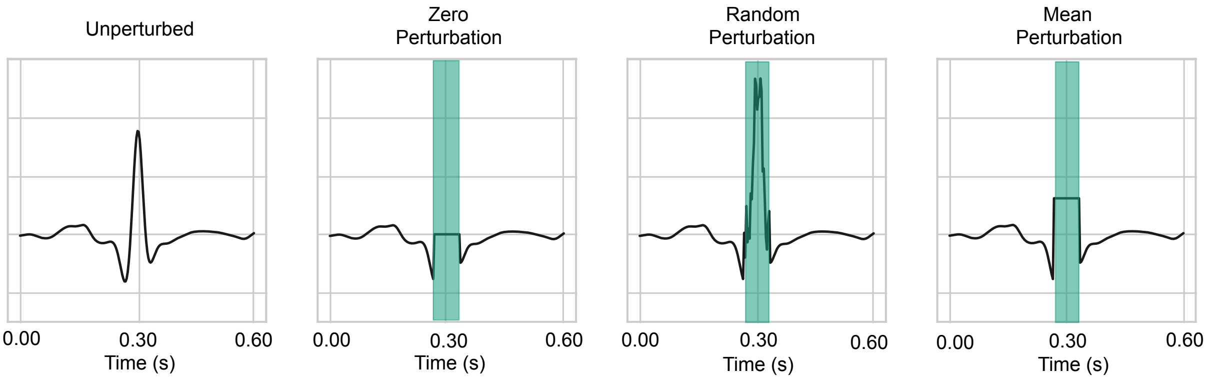


Figure 3: Example of the three possible perturbations, applied to the R peak of the heartbeat representation. On the left, the zero perturbation, followed by the random perturbation, and finally, on the right, the mean perturbation.

We propose three perturbation functions defined in Equations 7 to 9. Examples of those perturbations are illustrated on Figure 3.

The **zero** perturbation is defined as:

$$p(W_w) = 0 \tag{7}$$

The **random** perturbation is defined as:

$$p(W_w) = W_w + \theta \mathcal{N}(0, 1) \tag{8}$$

Where $\theta \in [0, 1]$ corresponds to a noise attenuation factor.

The **mean** perturbation consists of (I) calculating the average window value for all instances of D_{train}; and (II) averaging the calculated values of all instances of D_{train}:

$$p(W_w) = \frac{1}{Xl} \sum_{i=1}^{X} \sum_{j=1}^{l} X_i[(j-1)(l+1), jl] \tag{9}$$

3.2.3. Shapley Additive Explanations

SHAP is a method proposed by Slundberg et al. [23], that assigns importance values to each feature. These importance values are known as Shapley values and are obtained from the game theory.

Kernel SHAP is the model-agnostic approximation method and follows the same principle presented in Equation 3. Kernel SHAP combines LIME and Shapley values. It differs from LIME since it does not calculate the parameters heuristically, and assumes the local accuracy and consistency. The local accuracy, defined in Equation 10, assumes that the local approximation prediction matches the complex model's prediction. Consistency, Equation 11, assumes that perturbations applied are reversible:

$$\Omega(g) = 0 \tag{10}$$

$$\pi_{X_i} = \frac{(M-1)}{\binom{M}{|z|} |z| (M - |z|)} \tag{11}$$

where $\Omega(g)$ is a model complexity measurement, M is the number of simplified features, and z is the number of present samples in the perturbed dataset, samples with the value of 1 in the interpretable domain.

The assumed properties arrive at different parameters, which consequently lead to Shapley values calculation. To perturb an instance, kernel SHAP uses a mask function that contains vectors in the interpretable domain. This mask function attributes values of 1 or 0 to each sample of the perturbed instance. Similarly to LIME, if the value is 1 the sample remains unaltered. However, if the value is 0, the sample is replaced by the average value of the background dataset. To adapt Kernel SHAP into explaining windows of time series, the mask function was adapted, to force the absence (0) or presence (1) of all the values within a window.

3.3. Validation

To evaluate whether the sample's relevance in a time series is correctly calculated, we present two validation strategies: comparison with Shapelets using the Jaccard index and performance decrease.

The relevance weights for all windows are calculated using the methods presented in section 3.2. To validate the methods, the most relevant windows are replaced to evaluate the classifier's response. We consider as the most relevant windows the ones with relevance scores higher than a predefined threshold δ.

3.3.1. Jaccard Index

The Jaccard index or Jaccard similarity coefficient is a score that evaluates the similarity or diversity between two datasets. This metric measures correlation among two finite sets, A and B, and is represented as the size of their intersection divided by the size of the union, as defined:

$$J(A, B) = \frac{\#A \cap B}{\#A \cup B} \tag{12}$$

where # is the cardinality or number of elements in the set.

This method is often used in computer vision to evaluate image segmentation and object detection algorithms [44, 45]. Our purpose in using this similarity coefficient is to compare the most relevant windows calculated using the methods described in section 3.2 with shapelets extracted from the time series.

Shapelets are a time series primitive defined as the subsequences that can maximally describe a class [26]. This technique was developed to overcome some of the challenges of time series classification, namely the time and space complexity of instance-based classifiers. Shapelets can determine similarity based on smaller discriminative shapes, instead of using the entire time series length. The brute-force candidates search approach, based on an exhaustive search of shapelets candidates, suffers from high temporal complexity. A series of speed-up techniques were proposed, based on the early abandoning of distance computations and entropy pruning of the information gain metric [26]. The initial approach to leverage the advantage of using shapelets for time series classification was based on tree-based classifiers [26]. Thereafter, Lines et al. [46] suggested the shapelet transform, a space transformation before the classification, based on the minimum distances between the time series and the shapelet.

More recently, Grabocka et al. [47] propose an alternative approach based on learning shapelets. For such, the shapelets are firstly chosen randomly and are iteratively optimised to minimise the classification loss function. The learning process is based on generating subsequences that can linearly separate the distances from the dataset by their classes. Learning shapelets is based on shapelet transform. Subsequently, a linear learning model is used to make approximate predictions, as shown in Figure 4.

Since shapelets are subsequences that maximally describe a given class, we propose using the shapelet primitive to measure the similarity coefficient against the explanation methods for time series as outlined in Figure 5. The similarity is measured using the Jaccard index according to:

$$J(Shapelets, W_w) = \frac{\#Shapelets \cap W_w}{\#Shapelets \cup W_w} \tag{13}$$

J varies between 0 and 1. A value of 0 means that there was no match between the identified shapelets and the most relevant sequences. The value of 1 is the best-case scenario, where the most relevant windows are the same as the extracted shapelets.

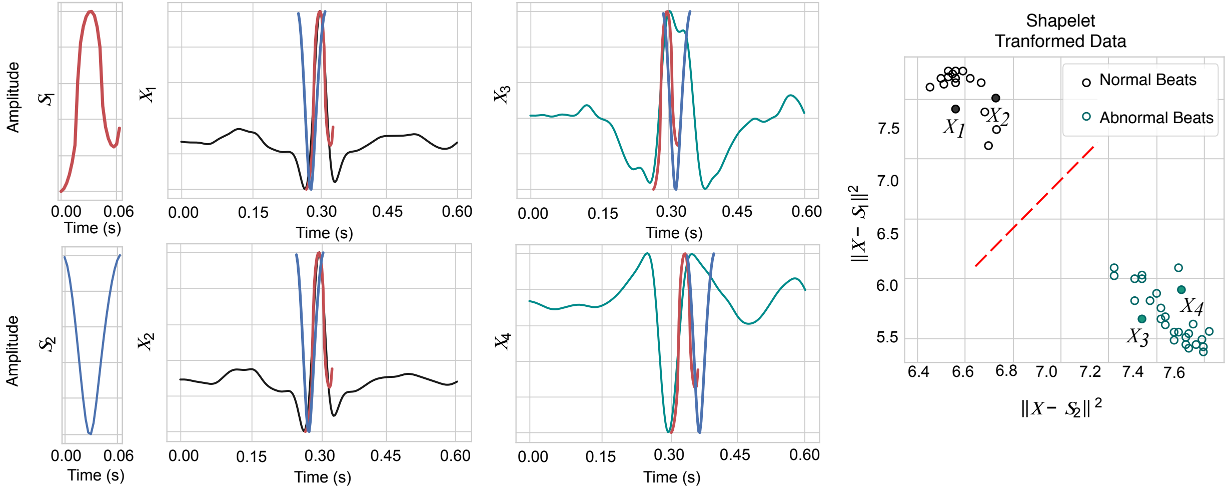


Figure 4: An illustration of two shapelets S1, S2 (leftmost plots) learned on the MIT-BIH ECG dataset. Time series' distances to shapelets can optimally project the series into a 2-dimensional space, entitled the shapelet-transformed representation [47] (rightmost plot). The middle plots show the intervals where the closest matches of the shapelets on the two series occurs.

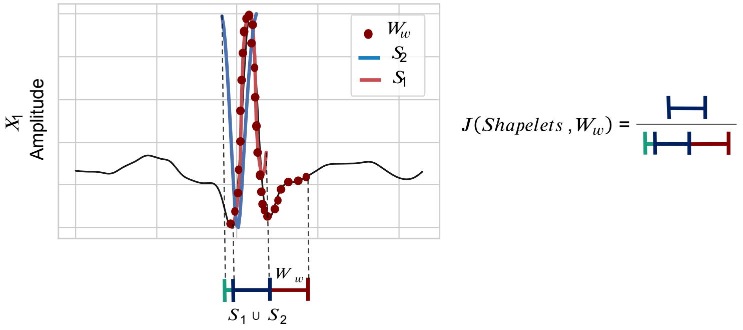


Figure 5: Example of Jaccard's index calculation trough the comparison of sets extracted from the shapelets classifier and the most relevant subsequences determined by the explanation method, where $Shapelets = S_1 \cup S_2$.

3.3.2. *Performance Decrease*

In this approach, the most relevant windows are replaced from the instance and the classifier's performance is recalculated. Different replacement methods are considered: zero, random and swap. The methods are applied to the windows W_w with relevance r_w equal to the maximum relevance score within the explanation, δ:

$$W'_w = \begin{cases} W_w, & \text{if } r_w < \delta \\ v(W_w), & \text{if } r_w = \delta \end{cases} \tag{14}$$

Where v is the replacement function.

The replacement methods are defined by Equations 15 to 18.

The **zero** substitution is defined as:

$$v(W_w) = 0 \tag{15}$$

The **random** substitutions is defined as:

$$v(W_w) = W_w + \theta \mathcal{N}(0, 1) \tag{16}$$

Where $\theta \in [0, 1]$ corresponds to a noise attenuation factor.

The **inverse** substitution is defined as:

$$v(W_w) = \max(X_i) - W_w \tag{17}$$

The **swap** substitution is defined as:

$$v(W_w) = \{x_{m+k}, x_{m+k-1} ..., x_m\} \tag{18}$$

The zero and random substitutions are similar to the perturbations outlined in Equations 7 and 8. The reverse and swap substitutions rely on symmetry about the amplitude and the time axis, respectively. The swap substitution is particularly relevant in the context of time series since it performs a replacement by the same subsequence in an opposite temporal ordering. If swapping the samples in their opposite temporal ordering produces the same explanation, it might imply that the sample's temporal dependency is not being taken into account.

3.4. Proposed Approach

A naïve approach to explain a time series consists of identifying the relevance of samples for the decision in the amplitude domain. However, the intrinsic nature of time series data often contains temporal dependencies between samples or among different time series. Explaining the samples' relevance using model-agnostic methods in the amplitude domain omit temporal dependencies, since the methods often assume sample independence and the amplitude only considers the Y-axis.

We propose that including the information of the time series derivative improves the quality of time series explanations. Including the derivative as a complement to improve time series classification has been proposed by Górecki and Łuczak [48, 49]. The derivative has also been used by Keogh and Pazzani [50] and Folgado et al. [51] to improve the alignment provided by the Dynamic Time Warping, a well-known similarity measurement in the time series domain. The aforementioned works reported in common that by using the derivative we are considering higher-level features, such as the shape of the time series. Therefore, the derivative has information that can be used by the explanation methods to integrate temporal dependency between samples.

The experimental protocol is summarised in Figure 6. Since our objective is to provide an empirical study of model-agnostic explanation methods for time series, we outlined a systematic approach composed of three steps - classify, explain and evaluate the quality of the explanations.

During the experiments, two different classifiers were considered. A k-Nearest Neighbour (k-NN) with $k = 5$ and a deep neural network. The classifiers were selected to investigate whether the different approaches assume different regions to be relevant for the classification.

The deep learning architecture differs in the binary and the multiclass case, due to better performance. The binary classifier is based on Kachuee et al. [52], composed of a five block Convolutional Neural Network (CNN): two convolution layers applying 1D convolution through time with variable kernels within 32, 64, and 128 of size 5; a 1-D max-pooling layer of size 5 and stride 2. The first convolutional layer has a kernel of 32, followed by two layers with 64 and the last two convolutional layers with 128. These blocks are followed by two fully-connected layers with 16 neurons and a softmax layer to predict output class probabilities. The multiclass classifier has four blocks with different parameters. The first convolutional layer has a kernel of 156, with a size of 27, and one stride. The following three layers have kernels of 64, 56, and 32, with sizes of 14, 3, and 1, respectively.

The models were trained with two combinations of input signals: (I) raw time series ($Amplitude$) and (II) combination of amplitude and derivative, composed of the concatenation of the two signals ($Amplitude + Derivative$).

The explanations are provided by calculating the relevance scores using the methods described in Section 3.2, namely PSI, LIME, and SHAP. The relevance scores were calculated from fixed-length windows of size $l = 24$

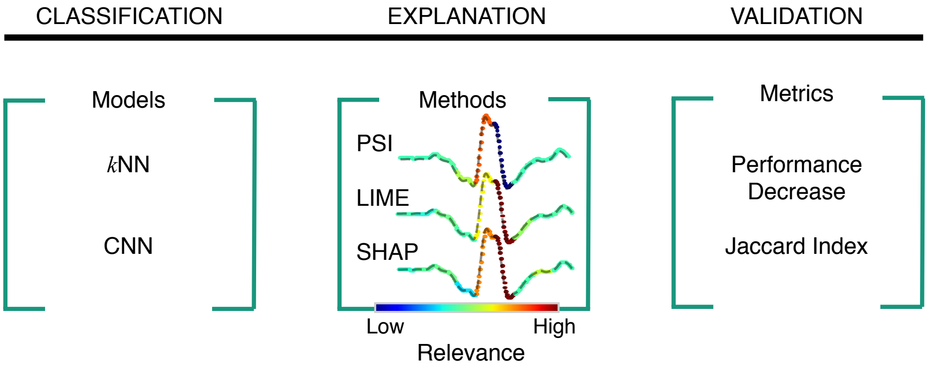


Figure 6: Schematic representation of the experimental protocol.

samples. The number of perturbed instances for each method was set to $k = 1000$. For the PSI, we performed three permutations for each slice (p).

We used the evaluation methods described in Section 3.3. Since the used XAI methods provide different distributions of relevance scores within the same instance, it is not advisable to use a threshold to define the most relevant windows, as it could imply considering a variable amount of windows into the validation. Thus, only one window is considered as the most relevant one.

The parameters used in this experiment were defined empirically and their impact on the results will be discussed in Section 5.

4. Results

In this section, we start by presenting the used dataset. The results are divided into binary and multiclass classification, presenting the results of the predictive performance with the algorithms and the validation of the explanations calculated using the proposed methods.

4.1. Dataset

The MIT-BIH Arrhythmia Database [53] is composed of 48 half-hour excerpts of two leads ambulatory ECG recordings from 47 subjects studied by the BIH Arrhythmia Laboratory. 23 ECG recordings were selected randomly in a set of 4000 24-hour ambulatory from a mixed population of inpatients and outpatients at Boston's Beth Israel Hospital. The remaining recordings were selected from the same set to include less common but clinically significant arrhythmias and assuring their presence in the dataset. Each beat has its R peak annotated and is classified, by two or more cardiologists, using 16 different labels. Table 3 shows how the classified beats are organised, particularly in normal (N) and ectopic classes (E), the latter further divided into ventricular ectopic beat (V), supra-ventricular ectopic beat (S), and fusion beat (F), according to the Association for the Advancement of Medical Information (AAMI) practices [54]. The dataset is split into train and test according to De Chazal et al. [55], which proposes a split that guarantees that the patients in the test set were not used during training. This split also discarded recordings from four patients with paced beats, as suggested by AAMI, remaining only 15 heartbeats of the Q class. Consequently, the Q class is ignored, resulting in a multiclass classification with four classes instead of five [56]. This dataset is highly unbalanced as it can be observed in Table 3, which makes the classification a challenging task. Further under-sampling was applied in order to adjust the class distribution of the considered dataset.

4.2. Classification

Since the dataset is highly unbalanced, the classifiers' performance in both the binary and multiclass cases does not report the accuracy, yet is assessed by measuring the recall, precision and F_1 score. Table 4 presents the models' performance for the binary classification. In this context, the two considered classes include the non-ectopic beats (N); and the set of four ectopic beats (E).

Table 4 shows evidence that the overall performance of the neural network is better when compared to the k-NN receiving the same input. When using the two signals concatenated, both the models do not perform that well, showing

Table 3
Class distribution of MIT-BIH arrhythmia database heartbeat types into the AAMI heartbeat classes.

AAMI heartbeat classes	MIT BIH-Arrythmia Database	Class Distribution
Non-Ectopic beats (N)	Normal beat Left bundle branch block beat Right bundle branch block beat Nodal escape beat Atrial escape beat	90330
Supraventricular ectopic beats (S)	Aberrated atrial premature beat Premature or ectopic supraventricular beat Atrial premature contraction Nodal escape beat	3024
Ventricular ectopic beats (V)	Ventricular flutter wave Ventricular escape beat Premature ventricular contraction	7235
Fusion beats (F)	Fusion of ventricular and normal beat	802

Table 4
Model's performance on a **binary** classification, classifying between non-ectopic (N) or ectopic (E) heartbeats. The F_1, recall and precision scores are presented in percentage (%).The best scores per metrics are highlighted in bold.

	F_1	Recall	Precision
k-NN $_{Amp}$	79.7	75.9	86.8
CNN$_{Amp}$	**90.3**	**89.5**	**91.9**
k-NN$_{Amp+Dev}$	71.8	65.6	83.8
CNN$_{Amp+Dev}$	73.2	67.2	86.4

Table 5
Model's performance on a **multiclass** classification, non-ectopic (N), and the ectopic heartbeats, including supraventricular (S), ventricular (V), and fusion (F). The F_1, recall and precision scores are presented in percentage (%). The best scores per metrics are highlighted in bold.

	F_1					Recall					Precision				
	N	S	V	F	Average	N	S	V	F	Average	N	S	V	F	Average
k-NN $_{Amp}$	74.7	7.6	62.2	0.1	70.5	62.2	23.7	**77.5**	0.8	61.1	93.6	4.5	52.0	0.0	86.5
CNN $_{Amp}$	**89.8**	**39.7**	**73.5**	**14.8**	**86.1**	**83.2**	**65.6**	71.9	**90.7**	**81.8**	**97.6**	**28.5**	**75.2**	**8.1**	**92.6**
k-NN $_{Amp+Dev}$	71.4	7.1	38.3	9.1	66.1	58.4	23.8	66.0	29.6	57.3	91.7	4.2	27.0	5.4	83.3
CNN $_{Amp+Dev}$	80.5	17.0	69.8	12.0	76.7	68.8	49.0	74.4	85.1	68.5	97.3	10.3	65.8	6.4	90.9

a decrease in the F_1, recall, and precision scores.

The multiclass case results are summarised in Table 5. The CNN approach is once again better performing than the k-NN. The average performance of the CNN$_{Amp}$ has higher values for F_1, recall and precision, when compared to the CNN$_{Amp+Dev}$.

Although the classifiers that only use the amplitude obtain better performances, the temporal information has clinical relevance and therefore in the following sections, we will evaluate the quality of the explanations with both approaches.

Table 6
Faithfulness measured by means of F$_1$, recall, precision scores and the mean R^2 (standard deviation), according to the different possible substitutions to produce the explanations for the binary classifications.

	k-NN $_{Amp}$			k-NN $_{Amp+Dev}$			CNN $_{Amp}$			CNN $_{Amp+Dev}$		
	Mean	Zero	Random	Mean	Zero	Random	Mean	Zero	Random	Mean	Zero	Random
F$_1$	91.9	**92.4**	89.1	95.2	93.1	**95.7**	**98.9**	97.1	98.5	**97.8**	95.3	87.8
Recall	73.6	76.8	**85.0**	**96.0**	91.1	94.3	97.6	**98.5**	97.7	95.3	96.2	**99.9**
Precision	**97.6**	95.5	77.2	91.0	89.7	**93.7**	**96.1**	85.6	93.3	**97.9**	94.4	71.7
R^2	0.61 (0.27)	**0.60 (0.20)**	0.27 (0.12)	**0.55 (0.22)**	0.47 (0.23)	0.20 (0.19)	0.59 (0.20)	**0.66 (0.16)**	0.57 (0.20)	**0.67 (0.17)**	0.56 (0.14)	0.40 (0.10)

Table 7
1-D Jaccard's index, measuring the similarity between shapelets and the most relevant subsequence identified by the explanation methods. The results are for the **binary** classification case.

	k-NN $_{Amp}$	k-NN $_{Amp+Dev}$	CNN $_{Amp}$	CNN $_{Amp+Dev}$
PSI	0.47 (0.32)	0.52 (0.32)	0.07 (0.13)	0.40 (0.35)
LIME	0.51 (0.31)	0.56 (0.30)	0.26 (0.33)	0.33 (0.36)
SHAP	0.22 (0.24)	0.01(0.08)	0.36 (0.33)	0.13 (0.26)
Random	0.11 (0.23)			

4.3. Explanation

The explanations are determined using the three methods described in Section 3.2. Additionally, the faithfulness of LIME was evaluated.

Faithfulness of LIME is the reliability of the local approximation to describe the complex classifier. Different explanations are created as a result of applying different substitutions in the time series. The reliability of approximation to the complex classifier will differ among different substitution methods. The fitting of the linear model to the complex model is represented through the R^2. Table 6 presents LIME's performance by comparing the predictions of its local approximation and the predictions from the complex classifier in the binary case.

The results suggest that the performance values, i.e. F$_1$, recall and precision are higher in most of the occasions for the mean and zero substitution. Nevertheless, the values for the Random substitution are also high. A most discriminative difference is measured by the R^2, since the R^2 values are consistently higher for the Mean and Zero in comparison with Random.

Both the mean and zero substitution methods present similar faithfulness. We chose the mean as the substitution method to be used in evaluating the quality of the explanations provided by LIME for the binary classification. To reduce computational time in the multiclass situation, it was assumed that the values of faithfulness were within the same range and the mean substitution was also chosen.

The results also indicate that the local approximation provided by LIME is adequately approximating the complex model.

4.4. Validation

The results to assess the quality of the explanations are introduced in the following sections, starting with the Jaccard index and then, performance decrease.

4.4.1. Jaccard Index

Jaccard's index is used as a metric to evaluate the explanation. As discussed in Section 3.3.1 it measures the similarity between the output from the shapelet-based classifier and the most relevant subsequences produced by the adapted XAI method. In the binary classification, the shapelet-based classifier had an F$_1$ score of 87.6%, a recall score of 86.8%, and precision of 88.7%.

Table 7 summarises the results for the three explanation methods considered for the considered classifiers. For baseline comparison, we included an additional explanation method we denoted as Random which assigns random relevance scores.

LIME and PSI show an increase in the Jaccard index when the derivative and amplitude are considered for both

Table 8
1-D Jaccard's index, measuring the similarity between shapelets and the most relevant subsequence identified by the explanation methods. The results are for the **multiclass** classification case.

	k-NN $_{Amp}$	k-NN $_{Amp+Dev}$	CNN $_{Amp}$	CNN $_{Amp+Dev}$
PSI	0.59 (0.46)	0.71 (0.42)	0.13 (0.30)	0.74 (0.40)
LIME	0.73 (0.41)	0.71 (0.42)	0.10 (0.27)	0.88 (0.27)
SHAP	0.18 (0.36)	0.10 (0.29)	0.44 (0.48)	0.05 (0.21)
Random	0.11 (0.23)			

Table 9
F_1 score's decrease of the **binary** situation.The F_1 score is measured after perturbing the most relevant window calculated for Random, PSI, LIME and SHAP. If the decrease is positive the F_1 score after the perturbation was lower than the initial score and negative otherwise.

	k-NN $_{Amp}$			**CNN $_{Amp}$**			**k-NN $_{Amp+Dev}$**			**CNN $_{Amp+Dev}$**		
	Zero	Inv	Swap	Zero	Inv	Swap	Zero	Inv	Swap	Zero	Inv	Swap
PSI	12.5	48.1	-4.2	61.0	80.2	4.3	22.3	32.4	29.3	20.1	44.8	-4.0
LIME	13.8	46.0	-2.5	37.9	35.1	1.7	27.8	30.6	31.8	24.0	25.1	-1.7
SHAP	1.1	60.6	-0.9	47.8	38.5	2.4	4.9	3.11	-0.2	4.4	49.3	-3.2
Random	0.8	34.7	-1.1	12.7	23.2	1.6	2.6	19.1	3.8	5.4	26.5	-0.6

the k-NN and CNN. The increase in PSI was more notable when comparing to LIME. SHAP does not behave well, showing scores below or on the range of the randomly attributed weights, amongst the different domains, for both Amp and $Amp + Dev$.

The results for the multiclass classification are presented in Table 8. In this case, the explanations are compared to a shapelet-based classifier with an F_1 score of 80.8%, recall of 74.9%, and precision of 88.7%.

In general, the Jaccard indexes are higher in the multiclass case in comparison to the binary case. The Jaccard index increases when considering the derivative, for k-NN$_{Amp+Dev}$ and CNN$_{Amp+Dev}$, with the explanations from PSI and LIME, respectively. The most similar explanations to the shapelets are the ones from LIME in the CNN$_{Amp+Dev}$, which not only presents the highest average score, but also the smallest standard deviation.

In contrast, the quality of the explanations measured by the Jaccard Index is lower for SHAP in comparison to PSI and LIME. Additionally, for SHAP, when the derivative is also considered the disagreement is even greater.

4.4.2. Performance Decrease

The previous section presented a high-level analysis regarding the agreement between the explanation methods and the shapelets, as a baseline method to retrieve the most relevant subsequences towards the classification. We present in this section a more detailed analysis that measures the quality of the explanations and tries to assert whether it is feasible to evaluate if the explanations take into account the temporal relationship between samples. Table 9 and 10 show the decrease in the F_1 score when the perturbations presented in Section 3.3.2 are applied, for the binary and the multiclass task, respectively.

Firstly, considering the results related to k-NN$_{Amp}$ and CNN$_{Amp}$, across all the explanation methods, it is reasonable to argue that the explanation is partly congruent with the model's behaviour. The perturbation of the most relevant subsequence led to an abrupt decrease in performance, in the case of the zero and inverse perturbations, albeit of lesser amplitude when compared to random weights. However, the performance decrease calculated using the swap perturbation did not change to a such extend. Since the swap perturbation modifies the temporal ordering of the samples without modifying their amplitude, one might indicate the temporal ordering of samples might not be relevant for the explanation method.

On the other hand, when evaluating the explanations for the models k-NN$_{Amp+Dev}$ and CNN$_{Amp+Dev}$, different outcomes arise. In the case of the k-NN $_{Amp+Dev}$, for PSI and LIME, the performance is more affected by swapping the samples' temporal ordering but is less affected by the inverse and zero perturbations, when compared to the random weights. Clarifying, one can say that the explanation for PSI and LIME, is less sensitive to the sample's amplitude and

Table 10
F_1 score's decrease of the **multiclass** situation.The F_1 score is measured after perturbing the most relevant window calculated for Random, PSI, LIME and SHAP. If the decrease is positive the F_1 score after the perturbation was lower than the initial score and negative otherwise.

	k-NN $_{Amp}$			CNN $_{Amp}$			k-NN $_{Amp+Dev}$			CNN $_{Amp+Dev}$		
	Zero	Inv	Swap	Zero	Inv	Swap	Zero	Inv	Swap	Zero	Inv	Swap
PSI	14.3	46.9	-4.6	33.0	53.1	29.5	23.4	23.3	20.5	52.6	30.7	10.6
LIME	18.2	46.7	-2.9	32.3	51.1	35.7	28.1	22.6	26.6	59.7	27.3	19.7
SHAP	-7.9	50.4	-5.4	5.3	17.4	4.2	5.2	10.4	6.4	8.6	59.9	3.1
Random	0.1	32.1	-1.4	3.6	20.4	5.0	3.7	18.5	3.3	3.4	56.7	1.1

more sensitive to its temporal ordering. Contrarily, SHAP presents values close or below the performance decrease of random relevance scores.

In the case of $\text{CNN}_{Amp+Dev}$, although there is a considerable loss of performance when applying the zero and inverse perturbations in the most relevant window, the same is not observed for the swap perturbation. Further discussion on this matter will be conducted in Section 5.

Similarly to the results for the binary classification for the k-NN_{Amp} there is a negligible variation of performance decrease for the swap perturbation in comparison to zero and inverse. However, it is worth to mention that the swap perturbation for CNN_{Amp} also yields to an abrupt drop in model performance, which will be further discussed in Section 5.

In general, the explanations provided by PSI and LIME are more sensitive to the temporal ordering when the derivative is also considered. This fact was not observed in SHAP, however, it's overall performance across all the methods used to validate the explanations was lower than PSI and LIME. Unlike the results for the binary case, the explanations for CNN predictions also show improvements in terms of temporal dependency with the inclusion of the derivative.

5. Discussion

The results presented in the last section support that both PSI and LIME are adequate methods to explain a time series classifier by measuring the relevance of each sample for the classification. These findings have a broad impact with regards to the applicability of such methods in real-world practice. For debugging machine learning model, a first application is, for instance, further understanding of why a model is misclassifying instances. Figure 7, is an example of that, presenting some of the false negatives for the S class, i.e., supraventricular ectopic beats that were classified as normal. The most relevant samples to make the predictions, in both cases, are located in the QRS segment.

The explanations suggest that similarly shaped beats have relevant samples in the same regions, only differing in the time that the QRS peak occurs. The correct classifications have their QRS peak occurring slightly earlier than the incorrect ones. It is verified that the further shifting of the misclassified heartbeats improves of the model's performance and produces the correct classifications.

Next, we further discuss the impact of the parameters which were defined empirically. In the window size, there is a trade-off between the slice's length and needed perturbed instances to generate a relevant explanation. To explain an instance using windows that are too short, the perturbed dataset must be sufficiently large, so that, in a random process of removing a window, it is possible to represent an opposite class or classes and thus have a relevance score associated. For applications where an increase in the temporal resolution of the explanation requires a reduced window size, it might constitute a challenge in terms of computational time complexity. The computational time for such an extent dataset might become untractable. On the other hand, windows that are too large end up not providing useful information about the shorter segments as they would end to contain distinct intervals that we might be interested in evaluating their relevance separately. Initial experiments were conducted and suggested that dividing the heartbeats into nine windows and using 1000 perturbed instances would be an adequate trade-off.

Regarding the different LIME perturbations to explain the predictions of ECG classifiers, faithfulness presents similar F_1 results for all the assessed perturbations. However, for the Random perturbation, R^2 results are somewhat lower which indicate that the linear regressions from LIME produce perturbations that are the least fitted to the classification when compared to the Mean and Zero.

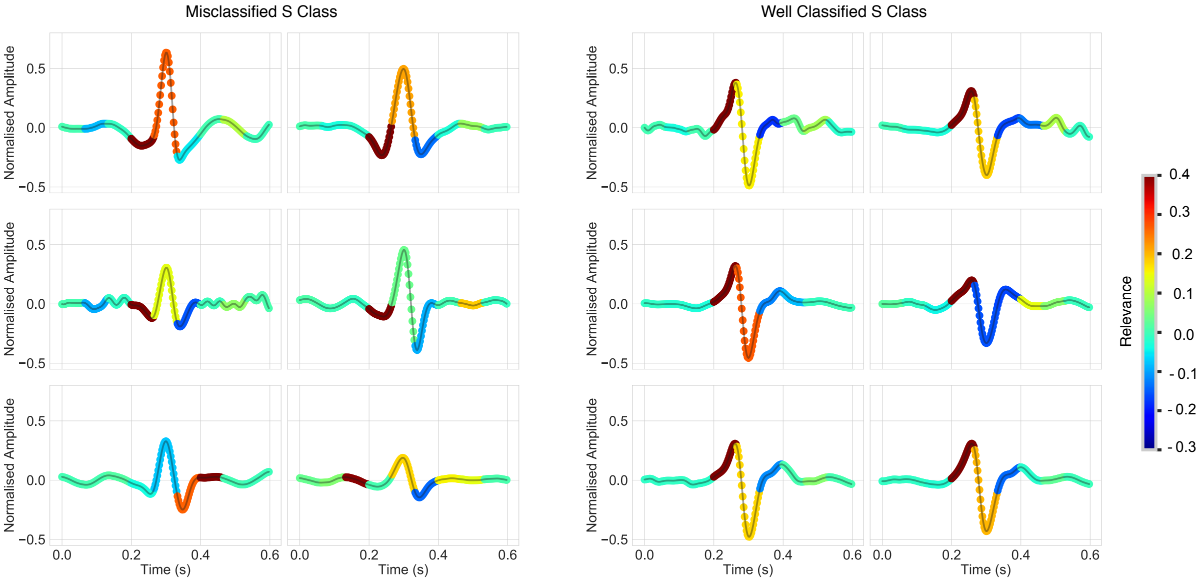


Figure 7: Representation of the CNN$_{Amp}$ misbehaviour. LIME explanations for false negatives of class S (on the left), and explanations for correct classifications of class S (on the right).

In the binary classification, the Jaccard index across the shapelet-based model and the explanations has values lower than 0.5, showing a low overlap between the shapelets and the most relevant window given by the XAI method. The high values of standard deviation indicate a variable agreement between the most relevant segments provided by the explanation and the model. Nevertheless, since only one segment was used to build the shapelet-based model, these values are in agreement with the work of Schlegel et al. [25], which reported identical low overlap between the shapelets and the XAI method when using less than two shapelets. In this case, the Jaccard's index is being deployed by assuming the shapelets classifier as a ground truth. However, that is far from reality as this classifier reports a specific performance and thus has an associated error. The 1-D Jaccard's index seems to be a reasonable metric to evaluate explanations, its particular use in time series explanations, still raises the question of which method could be used for comparison as the ground truth.

The multiclass classification has higher values for the Jaccard index. The increase in the similarity between the location of the shapelets and the most relevant window is due to the existence of less variability and more specificity amongst the time series of the same class. In the binary case the class the abnormal class, which includes different ectopic beats with characteristics that differ from each other, and thus it is more complex for a local explanation to be compatible with the behaviour of the shapelet model. In this case, the Jaccard index increases when considering the derivative, with the explanations from PSI and LIME, respectively.

In both situations, the explanations yielded by SHAP showed inadequate quality as the results were similar to the obtained with random weights.

With regards to the performance decrease, the results for LIME and PSI, using the k-NN, show that adding the derivative brings temporal information into the explanation.

In the binary classification, the CNN with the derivative does not show improvements in explaining the temporal dependency of samples, which may be related to the high variability of the abnormal class. For the multiclass case, the agnostic models convert the problem into binary and return the relevance of each window into deciding on the class identified by the complex classifier. Thus, each explanation generated is binary, in the sense that positive scores represent the positive contribution in such decision and negative scores represent the contribution of these windows to all the remaining classes. Therefore, the explanation, when consistent with the classifier, is more specific and has greater quality.

In the literature, SHAP is reported a reliable agnostic method for explaining predictions [21, 34]. However, in this particular case of ECG explanation, it did not obtain an adequate performance when compared to LIME and PSI, which indicates that our proposition is not feasible for SHAP. The low performance of SHAP was related to attributing

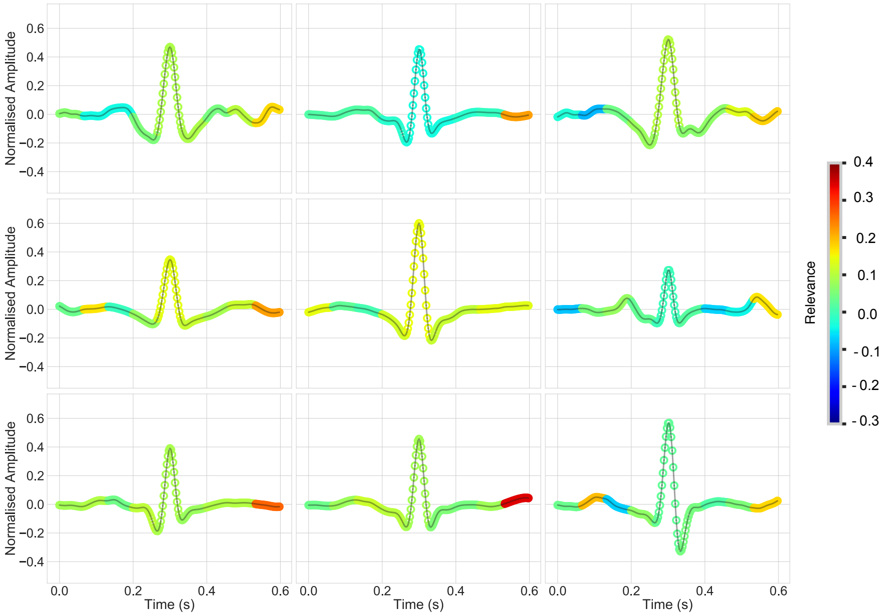


Figure 8: Representation of SHAP explanations for several predictions of the binary CNN$_{Amp+Dev}$.

high relevance scores to the last window as shown in Figure 8, in which are presented various explanations for correct classifications of normal beats.

Often the last window has greater relevance both, which produces explanations that are not consistent with the model. Therefore, both the Jaccard index and the performance decrease present results below the envisioned.

6. User study on ECG visual explanations

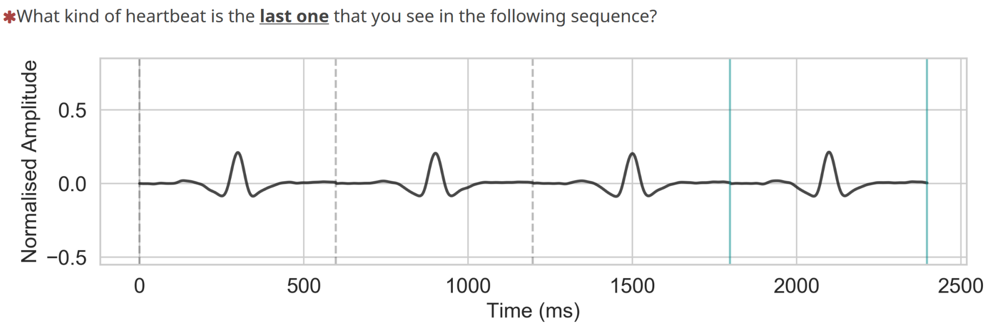
To evaluate the potential of the visual explanations described above to serve as sound interpretation aids, we designed a small user study in which three ECG readers were invited to read 20 ECGs, first without the visual explanation and then with this aid, and report their perceptions and comments on the support.

6.1. Method

To this aim, we developed an online questionnaire platform that could present, on random order, 20 ECGs, one case in each page (see Figure 10), encompassing 4 normalized heartbeats, of which only the last one had to be classified (see Figure 9). These ECGS were previously extracted by the reference dataset (MIT BIH Dataset) in order to equally represent its heterogeneity, and hence the sample encompassed 5 non-ectopic (normal) heartbeats and 15 pathological heartbeats, that is 5 supraventricular ectopic ones, 5 ventricular ectopic ones, and 5 fusion beats.

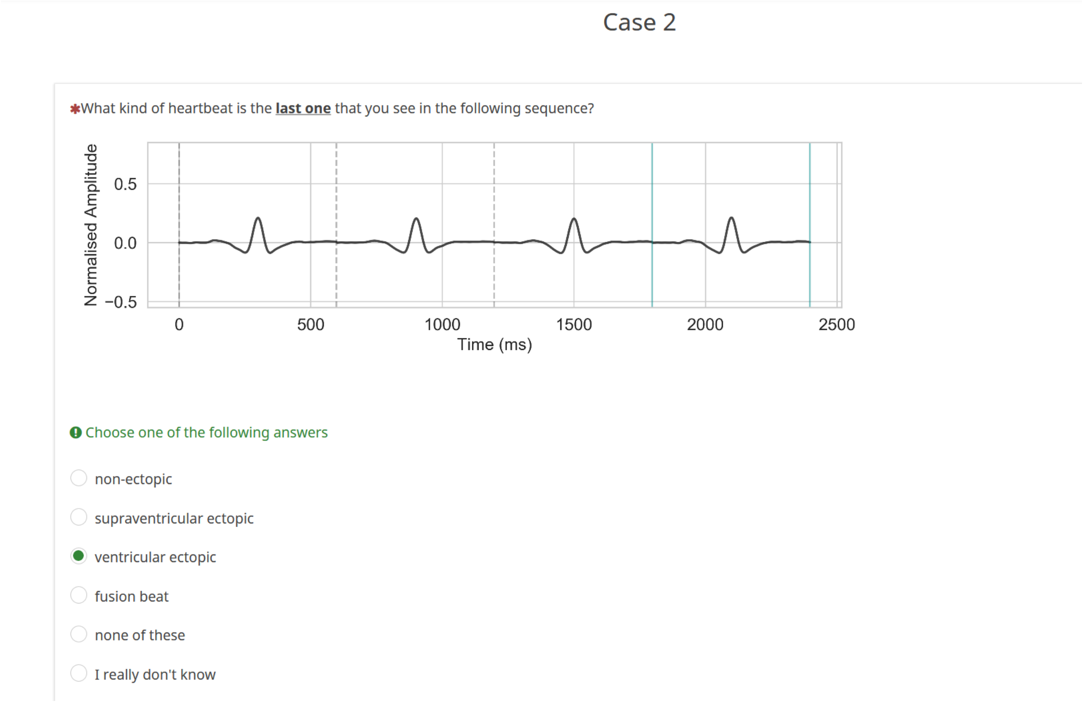
We then invited three ECG readers to partake our study, chosen both on a convenience basis and to have users representative of the widest range of potential ECG readers: one was a very knowledgeable cardiologist with more than 15 years of experience in reading ECGs; one (graduate) was a recent medicine graduate student who scored full marks in the electrocardiograph clinic class; one (resident) a resident who is specialising in geriatrics, which is a medical speciality where doctors are often called to prescribe, read and report ECG exams.

Each ECG was first presented without visual explanation and the participants were supposed to classify it choosing one of the following categories (see Figure 10: 'non-ectopic', 'supraventricular ectopic', 'ventricular ectopic', 'fusion beat', 'none of these' and 'I really don't know'. After this answer had been collected, the platform displayed the visual explanation for the heartbeat at hand, and the respondents could either confirm their previous classification or select a new one (see Figure 11, top side). Distinguishing between ECG reading before giving the visual explanation and after having given it allows to compute a pre-support accuracy and a post-support one, to see if the visual aid helps increase the readers' accuracy. Moreover, if the respondent selects a different class after seeing the explanation, we



✱What kind of heartbeat is the **last one** that you see in the following sequence?

Figure 9: An ECG taken from the experimental platform.



Case 2

✱What kind of heartbeat is the **last one** that you see in the following sequence?

Choose one of the following answers

○ non-ectopic

○ supraventricular ectopic

● ventricular ectopic

○ fusion beat

○ none of these

○ I really don't know

Figure 10: A screenshot from the experimental platform: how the visual explanation for a single ECG was presented and the available response options.

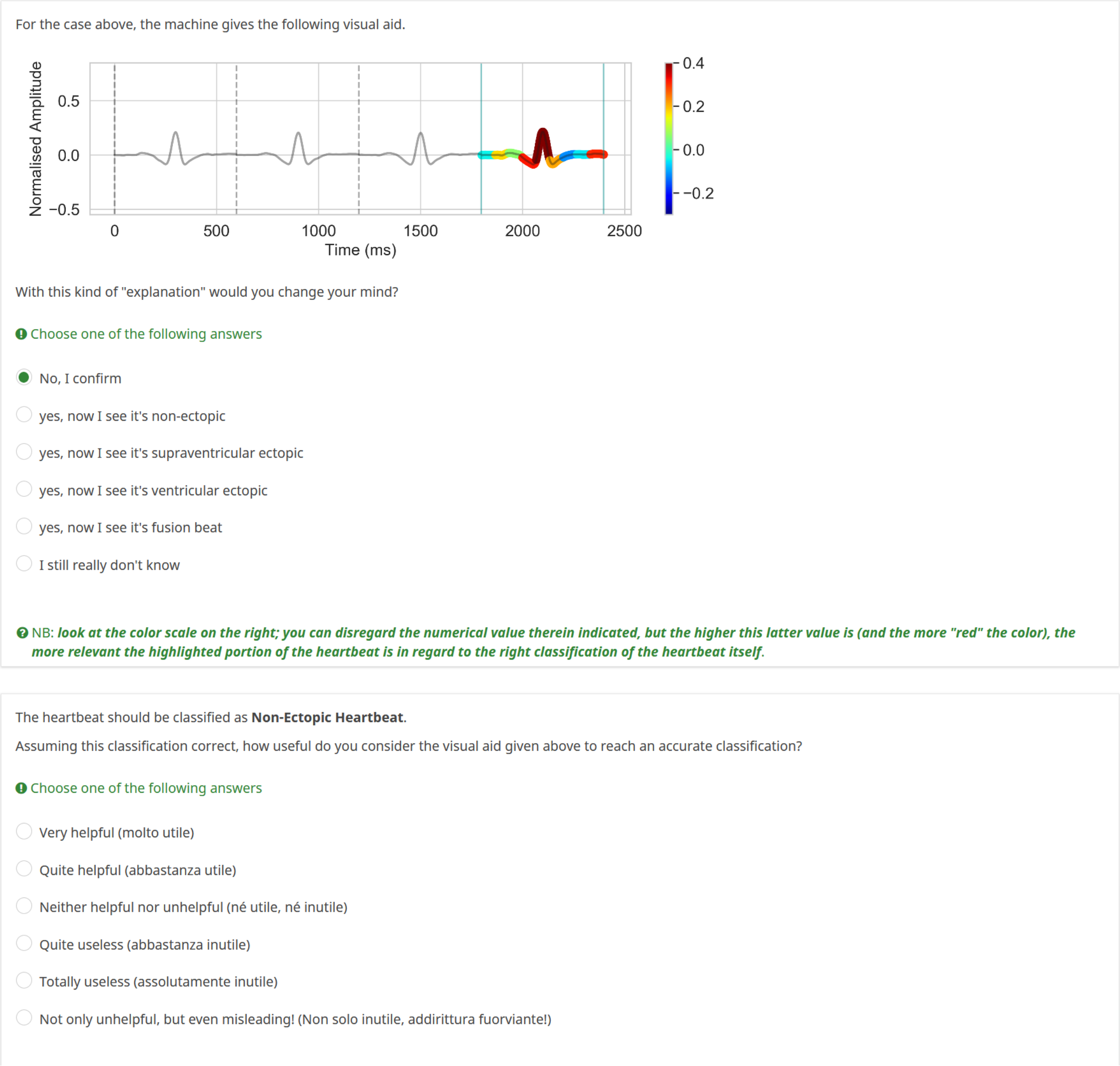
would record the event that the visual explanation has been effective in affecting the doctor's interpretation, and had them change their minds because of the patterns highlighted by the visual explanation.

After inserting the final diagnosis, the platform would display the right classification according to the gold standard and the respondents were invited to assess the usefulness of the visual explanation for the ECG at hand and the *typicalness* (see Figure 11, bottom side), that is the extent the visual explanation appeared *characteristic* of the heartbeat typology, at a post-hoc analysis. The respondent could also add a free-text comment.

At the end of the questionnaire, the platform asked the respondents to report about the overall perceived usefulness of the visual explanations, both in diagnostic tasks and for novice training, and to leave a general comment if they felt it necessary.

6.2. Results

The pre-support accuracy of the expert cardiologist was 70%; for the graduate it was 80%, and 65% for the resident. The cardiologist's performance suggests that the task was not trivial and that the selected heartbeats were representative



For the case above, the machine gives the following visual aid.

With this kind of "explanation" would you change your mind?

Choose one of the following answers

● No, I confirm

○ yes, now I see it's non-ectopic

○ yes, now I see it's supraventricular ectopic

○ yes, now I see it's ventricular ectopic

○ yes, now I see it's fusion beat

○ I still really don't know

NB: ***look at the color scale on the right; you can disregard the numerical value therein indicated, but the higher this latter value is (and the more "red" the color), the more relevant the highlighted portion of the heartbeat is in regard to the right classification of the heartbeat itself.***

The heartbeat should be classified as **Non-Ectopic Heartbeat**.

Assuming this classification correct, how useful do you consider the visual aid given above to reach an accurate classification?

Choose one of the following answers

○ Very helpful (molto utile)

○ Quite helpful (abbastanza utile)

○ Neither helpful nor unhelpful (né utile, né inutile)

○ Quite useless (abbastanza inutile)

○ Totally useless (assolutamente inutile)

○ Not only unhelpful, but even misleading! (Non solo inutile, addirittura fuorviante!)

Figure 11: A screenshot from the experimental platform: how the visual explanation for a single ECG was presented and the available response options.

of real-world cases and of the related complexity.

The difference in performance, although not significant, may be explained by the different attitude towards the task, as mirrored by the completion times (see Figure 12): the cardiologist considered each ECG for slightly more than 1 minute (69 s) and completed the whole task in approximately half an hour; the resident considered each ECG for slightly more than 2 minutes and half (157 s) and completed the task in slightly less than 1 hour (not necessarily consecutively); this difference is not statistically significant (t=1.71, df=38, p=0.097). However, the graduate took approximately 2 hours to complete the whole task of reading 20 ECGs. Differently from these two ECG readers, he left written comments for several ECGs, and admitted to have consulted some textbooks to interpret the ECGs as accurately as possible. In light of these considerations, the cardiologist's performance resembles cardiological practice more closely, as he also admitted to have mainly undertaken the interpretation of each ECG intuitively at a glance, as he would do during a real consultation. Conversely, the graduate took the experiment as an exercise where accuracy was a priority, in spite of unacceptable timing for real examinations. Finally, the resident performance represents a compromise between the former two different approaches, thus giving the experiment some external validity, albeit

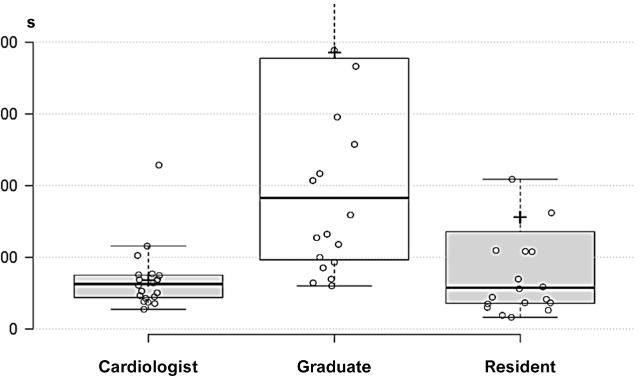


Figure 12: Boxplots of the completion times (in seconds) for each reader in the ECG interpretation task. Each circle corresponds to an ECG exam.

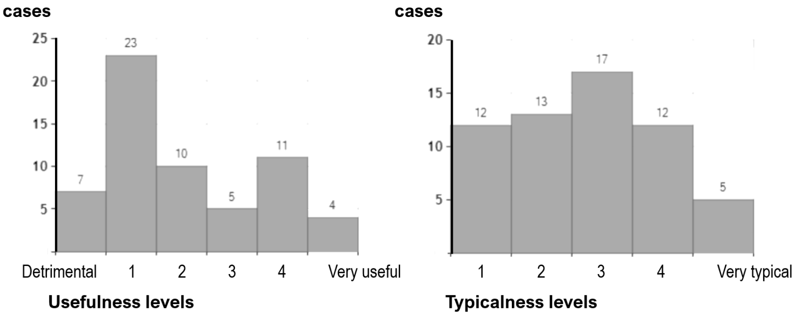


Figure 13: Histograms of the perceived usefulness and typicalness of the explanations used in the empirical study.

within the limits of a simulated exercise in a protected environment.

After seeing the visual explanations, the cardiologist and the resident worsened their performance by 5% (1 case). The decrease in accuracy was due to the fact that the former reader changed his mind for one case after seeing the visual explanation, and this latter misled him because he changed a right answer (normal beat) to a wrong one (abnormal one), thus inducing a false positive. Moreover, the visual explanation did not help the cardiologist avoid 6 errors, as he confirmed the initial wrong answer for 6 cases.

On the other hand, the graduate reader changed his answer two times after seeing the visual explanation, but only in one case this visual aid helped him choosing the correct diagnosis; in no case the system misled him with respect to to a correct answer. However, the system failed to help him avoid 4 errors, when he confirmed the wrong class for 4 cases. Finally, the resident changed his answers two times after seeing the visual explanation, and in no case the visual aid helped him choose the correct diagnosis; in fact, in one case the system misled him with respect to the correct answer (abnormal), inducing thus a false negative. Moreover, the provided explanations did not help the reader avoid 7 errors.

The user experience was mirrored by the overall perceived usefulness, which was low for all readers (2 on a 5-point scale), as well as in regard to the average potential they found for training (2). However, it is noteworthy that the perceived usefulness of the single explanations was more various (see Figure 13, on the left, and Figure 14). On a scale from 0 (detrimental) to 5 (very useful) the average score by the expert ECG reader was 1.6 (SD=0.9) with one case that was associated with 0 (the case in which the cardiologist was misled) and one case that was associated with 4. The average score of the resident was similar (m=1.4, SD=1.4), with 5 cases for which the visual explanation was considered potentially misleading, and 3 cases that conversely had been associated with an aid that was found quite useful (4). Conversely, the graduate reader perceived a significantly higher usefulness (m=3.1, SD=1.6), with only 1 case associated with a detrimental explanation and 4 cases with a very useful support (5), especially for abnormal ECGs. Notably, the typicalness of the explanations was perceived to significantly be higher (see Figure 13, on the right

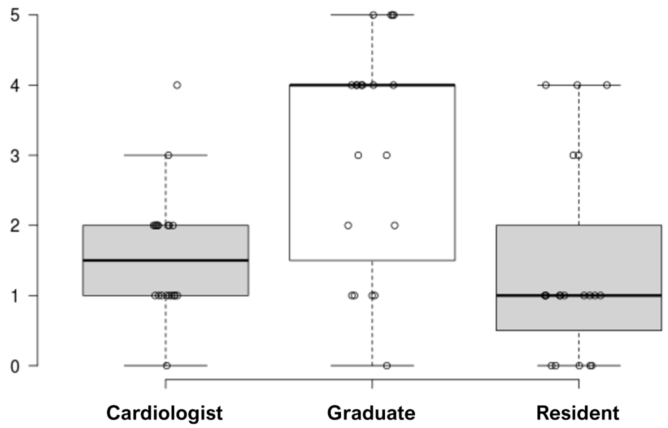


Figure 14: Boxplots of the perceived usefulness of the visual explanations reported on a 0-to-5 scale by the participants of the ECG interpretation task. Circles correspond to the explanations of the single ECGs. We recall that explanations associated with zero were considered detrimental and potentially misleading the ECG interpretation.

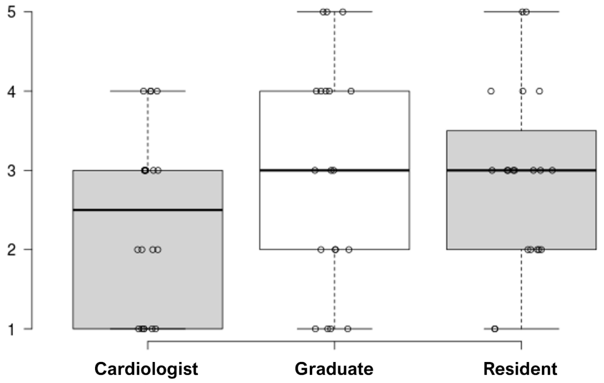


Figure 15: Boxplots of the perceived typicalness of the visual explanations reported on a 1-to-5 scale by the participants of the ECG interpretation task. Circles correspond to the explanations of the single ECGs.

and Figure 15): on average 2.75 (SD=1.24) on a 1 to 5 scale with small differences among the readers. Moreover, we detected a statistically significant difference (H=9.12, N = 59, p-value = .028) in regard to the perceived typicalness across the different kinds of heartbeat (see Figure 16): the visual explanations were perceived characteristic for abnormal beats, especially for ventricular ectopic ones (median typicalness: 4): this suggests that the employed method could be appreciated especially by less expert readers, and in training settings to have students observe the most salient parts of the signal for recognising ventricular ectopic heartbeats.

6.3. Discussion

For the intrinsically qualitative nature of the above user study, in what follows we interpret the grim figures reported above in the light of the comments reported by the respondents during the task.

The expert cardiologist motivated his low mark for the usefulness of the visual explanation in light of the fact that the interpretation of ECG beats is primarily a rule-based process, that follows clear-defined heuristics and is usually performed consulting much longer strips on graph paper and, generally, clinical information regarding the case (e.g., patient history, current therapies, previous exams). As said above, the graduate felt strongly engaged in the experiment and provided several comments during its execution: although he considered some visual explanations very useful, he also agreed upon the relative low usefulness of this kind of support in clinical practice. He motivated this judgement by noticing that the machine seems to focus on elements that the traditional textbooks of cardiology does not consider

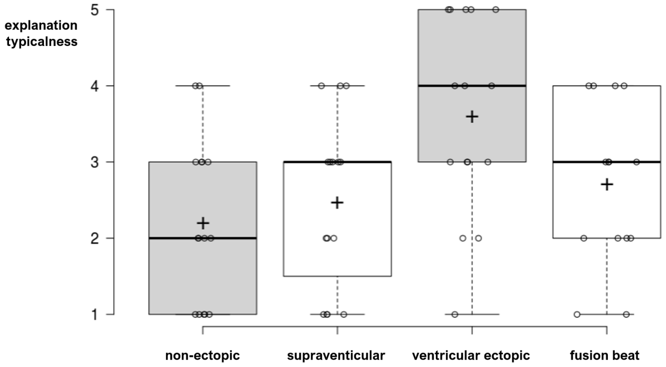


Figure 16: Boxplots of the perceived typicalness of the visual explanations grouped by kind of heartbeat. Circles correspond to the explanations of the single ECGs. Averages are indicated with a small cross.

and seems to neglect the pivotal element that have been codified into the cardiography rules mentioned also by the cardiologist above: "ECG readers are specifically trained to look for different clues in ECG tracings than the shapes highlighted by the AI" the graduate said. For example, considering the fusion beats, a clinician usually refers to criteria first proposed in [57]: these rules focus on the duration of the P-R and P-S parts of the fusion beat compared with the duration of the corresponding parts of the supraventricular heartbeat. Instead, the visual explanation neglected these characteristics, and drew attention to other deflections or intervals. This suggests a provocative conjecture: the visual explanations regards aspects of the heartbeat that are characteristic of a certain pathological condition that are imperceptible to the human eye (unless properly highlighted): their study could lead to a new, and possibly more efficient and effective, approach to ECG interpretation, although that would impossible without the technological support. Indeed, the graduate also noticed that once a user gets acquainted with the seemingly odd ways in which the machine uses the coloured patterns to explain its decision, looking directly at the visual explanation superimposed on the heartbeat would seem a more convenient and effective method than looking at the signal and explanation separately. However, he was also aware that doing so could make the reader (especially if rookie) overdependent on the visual aid and less prone to reflection.

The comments above suggests the following distinction. Explanations can be divided according to whether: a) they are aimed at helping users understand *why they should believe the given answer is true*; b) they are aimed at helping users understand *why the system has proposed that answer and not others*. These are ontologically different kinds of explanations: The first kind of explanation can be denoted as *epistemic* (or persuasive), because it regards properties of the phenomenon of interest or the related causal, explanatory mechanisms; the second kind can be seen as *gnoseological* or justificative, as it regards the inner ways by which a computational model discriminates the phenomenon in some way, or what was relevant for its final output.

Therefore, our small empirical study, despite being small for both the respondent and ECG sample, suggests that: superimposing sample relevance on a ECG time series acts more as a *gnoseological explanation*, which hints at the implicit (and mostly morphological) criteria by which the model yields its predictions, than as an epistemic explanation giving cardiologist clues on what to observe in the heartbeat for its right classification. However, as noticed by the graduate participant, training doctors-to-be with this system could make them able to understand the morphological correlates between the emphasised shapes and the right classification and exploit these "explanations" as complementary indications with respect to the established electrocardiographic rules and other available clinical information.

7. Conclusions

The lack of interpretability of top-line performance models has hindered the acceptance of AI in the medical field . Moreover, there is still a lack of consensus about how to assess the quality and usefulness of explanations given by these models [12]. Furthermore, model-agnostic automatic explanation of time series classification still poses grand challenges and is at a preliminary stage of maturity, as the existing methods often assume feature independence which, when translated to the time series domain, would imply temporal independence between samples.

As previously mentioned, in this work we mainly focused on the first level of the taxonomy proposed, that is sample-based explanations, while the other three levels of explanations, the feature-, morphology- and text-based ones, require further future work. In particular, we believe that exploring the potential applications of shapelets, which we employed to evaluate the proposed sample-based explanation methods, could be of particular interest: as shapelets represent morphological characteristics of a (collection of) time series that are discriminative for a given class, the development of a XAI approach based on shapelet classifiers could be relevant for developing feature-based (or morphology-based) explanations methods. Another pertinent research line for future work could regard automatic segmentation, to define variable-length windows that properly explain segments on the basis of the signal morphology, instead of an arbitrarily fixed length. Finally, we mention that while in this work we relied on an implementation of agnostic explanation methods based on standard Lasso regression, exploring variations of the Lasso approach more suited to the considered domain, such as Group Lasso [58, 59] or Fused Lasso [60], could be an interesting research direction, so as to better account for the temporal dimension in the computed explanations.

Summing things up: in this paper we presented a practical taxonomy for time-series explanation, and focused on sample-based visual explanations. To address the above challenges, we adapted several model-agnostic XAI methods to detect and represent the relevance of each window for the classification of a given instance. Our main contributions are: an extensive evaluation of several explanation methods for time series in the MIT BIH Dataset; a method to create visual explanations which exploit the derivative of the time series, seen as a convenient way to introduce the notion of temporal dependency into the explanation (since the derivative is, by definition, the instantaneous rate of change of a signal); a technical validation of the method mentioned above with respect to other model-agnostic methods, by using a public ECG dataset; and finally, an empirical evaluation of the usefulness of the resulting visual explanations in a realistic experiment involving three ECG readers, where the potential of this technology was appraised as a complementary aid to heartbeat interpretation that does not substitute but rather supplement other traditional means to help cardiovascular experts and technicians in their daily practice.

Overall, this study aims to provide a basis for the development and validation of intuitive and reliable explanations that could be based on the proposed methodology and upon which to undertake further research in realistic scenarios and real-world settings.

Acknowledgments

The authors gratefully acknowledge the contribution of Raffaele Rasoini, M.D. and Paolo Blangiardo, M.D., for their time and participation in the experimental user study.

References

[1] World Health Organization, Cardiovascular diseases (CVDs), 2017. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] M. Sharma, R. S. Tan, U. R. Acharya, Automated heartbeat classification and detection of arrhythmia using optimal orthogonal wavelet filters, Informatics in Medicine Unlocked 16 (2019) 100221. URL: https://doi.org/10.1016/j.imu.2019.100221. doi:10.1016/j.imu.2019.100221.

[3] W. Fan, J. Liu, S. Zhu, P. M. Pardalos, Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS), Annals of Operations Research (2018) 1–26. doi:10.1007/s10479-018-2818-y.

[4] F. Cabitza, A. Campagner, C. Balsano, Bridging the "last mile" gap between ai implementation and operation:"data awareness" that matters, Annals of Translational Medicine 8 (2020).

[5] T. Sullivan, Half of hospitals to adopt artificial intelligence within 5 years | Healthcare IT News, 2017. URL: https://www.healthcareitnews.com/news/half-hospitals-adopt-artificial-intelligence-within-5-years.

[6] F. Cabitza, Biases affecting human decision making in ai-supported second opinion settings, in: International Conference on Modeling Decisions for Artificial Intelligence, Springer, 2019, pp. 283–294.

[7] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): towards medical XAI, CoRR abs/1907.07374 (2019). URL: http://arxiv.org/abs/1907.07374. arXiv:1907.07374.

[8] Z. C. Lipton, The mythos of model interpretability, Communications of the ACM 61 (2018) 35–43. doi:10.1145/3233231. arXiv:1606.03490.

[9] F. Cabitza, D. Ciucci, R. Rasoini, A giant with feet of clay: On the validity of the data that feed machine learning in medicine, in: Organizing for the digital world, Springer, 2019, pp. 121–136.

[10] F. Cabitza, A. Campagner, L. M. Sconfienza, As if sand were stone. new concepts and metrics to probe the ground on which to build trustable ai, BMC Medical Informatics and Decision Making 20 (2020) 1–21.

[11] B. Goodman, S. Flaxman, European union regulations on algorithmic decision making and a "right to explanation", AI Magazine 38 (2017) 50–57. doi:10.1609/aimag.v38i3.2741. arXiv:1606.08813.

[12] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, WIREs Data Mining and Knowledge Discovery 9 (2019). URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1312. doi:10.1002/widm.1312.

[13] F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, 2017. URL: http://arxiv.org/abs/1702.08608. arXiv:1702.08608.

[14] S. Liu, X. Wang, M. Liu, J. Zhu, Towards better analysis of machine learning models: A visual analytics perspective, Visual Informatics 1 (2017) 48–56. doi:10.1016/j.visinf.2017.01.006. arXiv:1702.01226.

[15] E. Shortliffe, Computer-based medical consultations: MYCIN, volume 2, Elsevier, 2012.

[16] M. S. Mahoney, The History of Computing in the History of Technology, IEEE Annals of the History of Computing 10 (1988).

[17] A. H. Gee, D. Garcia-Olano, J. Ghosh, D. Paydarfar, Explaining deep classification of time-series data with learned prototypes, CEUR Workshop Proceedings 2429 (2019) 15–22. arXiv:1904.08935.

[18] H. Song, D. Rajan, J. J. Thiagarajan, A. Spanias, Attend and diagnose: Clinical time series analysis using attention models, 32nd AAAI Conference on Artificial Intelligence, AAAI 2018 (2018) 4091–4098. arXiv:1711.03905.

[19] L. Lin, B. Xu, W. Wu, T. W. Richardson, E. A. Bernal, Medical time series classification with hierarchical attention-based temporal convolutional networks: A case study of myotonic dystrophy diagnosis, CoRR abs/1903.11748 (2019). URL: http://arxiv.org/abs/1903.11748. arXiv:1903.11748.

[20] F. Horst, S. Lapuschkin, W. Samek, K. R. Müller, W. I. Schöllhorn, Explaining the unique nature of individual gait patterns with deep learning, Scientific Reports 9 (2019). doi:10.1038/s41598-019-38748-8. arXiv:1808.04308.

[21] F. Mujkanovic, Explaining the Predictions of Any Time Series Classifier, Bachelor's thesis, Universität Potsdam, Potsdam, Germany, 2019.

[22] M. Guilleme, V. Masson, L. Roze, A. Termier, Agnostic local explanation for time series classification, Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI 2019-November (2019) 432–439. doi:10.1109/ICTAI.2019.00067.

[23] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[24] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?" Explaining the Predictions of Any Classifier, in: Proceedings of the 2016 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Demonstrations, Association for Computational Linguistics, 2016, pp. 97—-101. URL: http://dx.doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[25] H. Arnout, M. El-Assady, D. Oelke, D. A. Keim, Towards a rigorous evaluation of xai methods on time series, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, 2019, pp. 4197–4201.

[26] L. Ye, E. Keogh, Time series shapelets: A new primitive for data mining, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 947–956. URL: https://doi.org/10.1145/1557019.1557122. doi:10.1145/1557019.1557122.

[27] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE, 2018, pp. 80–89.

[28] A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[29] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115. doi:10.1016/j.inffus.2019.12.012.

[30] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, et al., One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, arXiv preprint arXiv:1909.03012 (2019).

[31] J. Rodrigues, D. Folgado, D. Belo, H. Gamboa, SSTS: A syntactic tool for pattern search on time series, Information Processing and Management 56 (2019) 61–76. URL: https://doi.org/10.1016/j.ipm.2018.09.001. doi:10.1016/j.ipm.2018.09.001.

[32] W. Gale, L. Oakden-Rayner, G. Carneiro, L. J. Palmer, A. P. Bradley, Producing radiologist-quality reports for interpretable deep learning., in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 1275–1279.

[33] J. Sarkar, C. Peterson, Enabling Prognostics of Robust Design with Interpretable Machine Learning, Technical Digest - International Electron Devices Meeting, IEDM 2019-December (2019) 286–289. doi:10.1109/IEDM19573.2019.8993481.

[34] C. Molnar, G. König, B. Bischl, G. Casalicchio, Model-agnostic feature importance and effects with dependent features–a conditional subgroup approach, arXiv preprint arXiv:2006.04628 (2020).

[35] C. Schockaert, V. Macher, A. Schmitz, Vae-lime: deep generative model based approach for local data-driven model interpretability applied to the ironmaking industry, arXiv preprint arXiv:2007.10256 (2020).

[36] L. Hu, J. Chen, V. N. Nair, A. Sudjianto, Locally interpretable models and effects based on supervised partitioning (lime-sup), arXiv preprint arXiv:1806.00663 (2018).

[37] R. Elshawi, Y. Sherif, M. Al-Mallah, S. Sakr, Interpretability in healthcare a comparative study of local machine learning interpretability techniques, Proceedings - IEEE Symposium on Computer-Based Medical Systems 2019-June (2019) 275–280. doi:10.1109/CBMS.2019.00065.

[38] Y. R. Xie, D. C. Castro, S. E. Bell, S. S. Rubakhin, J. V. Sweedler, Single-Cell Classification Using Mass Spectrometry through Interpretable Machine Learning, Analytical Chemistry (2020). doi:10.1021/acs.analchem.0c01660.

[39] L. Breiman, Random forests, Machine Learning 1 (2001) 5–32. doi:10.1201/9780367816377-11.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[41] C. Molnar, Interpretable Machine Learning, 2019. https://christophm.github.io/interpretable-ml-book/.

[42] P. Kopper, Lime and neighbourhood, in: C. Molnar (Ed.), Limitations of Interpretable Machine Learning Methods, 2019, pp. 201–222. URL: https://compstat-lmu.github.io/iml{_}methods{_}limitations/.

[43] T. Laugel, X. Renard, M.-J. Lesot, C. Marsala, M. Detyniecki, Defining locality for surrogates in post-hoc interpretablity, arXiv preprint arXiv:1806.07498 (2018).

[44] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, M. B. Blaschko, Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice, Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 (2019) 92–100. URL: http://dx.doi.org/10.1007/978-3-030-32245-8_11. doi:10.1007/978-3-030-32245-8_11.

[45] Y. Yuan, M. Chao, Y. Lo, Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance, IEEE Transactions on Medical Imaging 36 (2017) 1876–1886.

[46] J. Lines, L. M. Davis, J. Hills, A. Bagnall, A shapelet transform for time series classification, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2012) 289–297. doi:10.1145/2339530.2339579.

[47] J. Grabocka, N. Schilling, M. Wistuba, L. Schmidt-Thieme, Learning time-series shapelets, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 392–401. URL: https://doi.org/10.1145/2623330.2623613. doi:10.1145/2623330.2623613.

[48] T. Górecki, M. Łuczak, Using derivatives in time series classification, Data Mining and Knowledge Discovery 26 (2013) 310–331.

[49] T. Górecki, M. Łuczak, First and second derivatives in time series classification using dtw, Communications in Statistics-Simulation and Computation 43 (2014) 2081–2092.

[50] E. J. Keogh, M. J. Pazzani, Derivative dynamic time warping, in: Proceedings of the 2001 SIAM international conference on data mining, SIAM, 2001, pp. 1–11.

[51] D. Folgado, M. Barandas, R. Matias, R. Martins, M. Carvalho, H. Gamboa, Time alignment measurement for time series, Pattern Recognition 81 (2018) 268–279.

[52] M. Kachuee, S. Fazeli, M. Sarrafzadeh, ECG heartbeat classification: A deep transferable representation, CoRR abs/1805.00794 (2018). URL: http://arxiv.org/abs/1805.00794. arXiv:1805.00794.

[53] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals., Circulation 101 (2000). doi:10.1161/01.cir.101.23.e215.

[54] A. for the Advancement of Medical Instrumentation, A. N. S. Institute, Testing and Reporting Performance Results of Cardiac Rhythm and ST-segment Measurement Algorithms, ANSI/AAMI, The Association, 1998.

[55] P. De Chazal, M. O'Dwyer, R. B. Reilly, Automatic classification of heartbeats using ECG morphology and heartbeat interval features, IEEE Transactions on Biomedical Engineering 51 (2004) 1196–1206. URL: https://pubmed.ncbi.nlm.nih.gov/15248536/. doi:10.1109/TBME.2004.827359.

[56] V. Mondéjar-Guerra, J. Novo, J. Rouco, M. G. Penedo, M. Ortega, Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers, Biomedical Signal Processing and Control 47 (2019) 41–48. doi:10.1016/j.bspc.2018.08.007.

[57] H. J. MARRIOTT, N. L. SCHWARTZ, H. H. BIX, Ventricular fusion beats, Circulation 26 (1962) 880–884.

[58] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (2006) 49–67.

[59] J. Friedman, T. Hastie, R. Tibshirani, A note on the group lasso and a sparse group lasso, arXiv preprint arXiv:1001.0736 (2010).

[60] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2005) 91–108.