Maria Inês Lourenço das Neves

Bachelor of Science in Biomedical Engineering

Opening the black-box of artificial intelligence predictions on clinical decision support systems

Dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Science in
Biomedical Engineering

Adviser: Prof. Hugo Filipe Silveira Gamboa, Assistant Professor,
NOVA University of Lisbon

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

November, 2020

Opening the black-box of artificial intelligence predictions on clinical decision support systems

This document was created using the (pdf)LaTeX processor, based on the "novathesis" template[1], developed at the Dep. Informática of FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Acknowledgements

The conclusion of this stage was only possible with the help of several people who stayed by my side even in the most difficult moments.

First of all, I would like to express my gratitude to my academic supervisor, Professor Hugo Gamboa, who gave me the opportunity to explore such an interesting and stimulating topic, always providing me with the bases to develop it. I would also like to thank the entire *Fraunhofer AICOS* members, in particular to the Lisbon office, for welcoming me and remaining close even in the odd pandemic context. To Duarte Folgado, I could never thank enough, for all its patience and extremely wise advise. I hope I carry some of your knowledge and organisation skills. Thanks also to Sara Santos for guiding me, not only around Lisbon but also throughout this journey.

To my college friends, I express my sincere gratitude, for making this last five years such a fun experience. A special thank to my deepest friends, Beatriz, Madalena, Joana, Miguel, Lourenço and João that are the best thing that the university has given me, and without whom I could never have accomplished this.

To my longtime friends, João Sebastião, Manuel and Bárbara, with whom I share an undeniable connection, a big thank you for helping me see the world in a relaxed and positive way.

Finally, I am extremely grateful for my family. To my sister for always setting an example and believing in me. To my brother for the companionship and constantly testing the limits of my patience. To José Álvaro for never forgetting a special event and for being a language expert. To my mother, whom I owe everything, I am left beyond words. I am eternally grateful for the unconditional love and for the values you have taught me. I aspire to be like you.

ECG

Recorre hoje à ciência Atingirás
talvez o impossível explicar
um coração humano os seus inúteis
caprichos
a horas inverosímeis os maiores
terrores todas as suas
tempestades

Imagina de novo se quiseres
um electrocardiograma
e tenta prescutar no seu traçado
o ritmo sinusal das emoções
o assombro de um transe numa breve
extra-sístole
a acalmia de cada onda T
ou esse tempo demasiado longo
entre as suas pequenas ondas P
e o seu QRS isso a que chamas
um bloqueio de primeiro grau

É inofensivo Não te preocupes
Não está bloqueado esse
coração

Fernando Pinto do Amaral

Abstract

Cardiovascular diseases are the leading global death cause. Their treatment and prevention rely on electrocardiogram interpretation, which is dependent on the physician's variability. Subjectiveness is intrinsic to electrocardiogram interpretation and hence, prone to errors. To assist physicians in making precise and thoughtful decisions, artificial intelligence is being deployed to develop models that can interpret extent datasets and provide accurate decisions. However, the lack of interpretability of most machine learning models stands as one of the drawbacks of their deployment, particularly in the medical domain. Furthermore, most of the currently deployed explainable artificial intelligence methods assume independence between features, which means temporal independence when dealing with time series. The inherent characteristic of time series cannot be ignored as it carries importance for the human decision making process.

This dissertation focuses on the explanation of heartbeat classification using several adaptations of state-of-the-art model-agnostic methods, to locally explain time series classification. To address the explanation of time series classifiers, a preliminary conceptual framework is proposed, and the use of the derivative is suggested as a complement to add temporal dependency between samples. The results were validated on an extent public dataset, through the 1-D Jaccard's index, which consists of the comparison of the subsequences extracted from an interpretable model and the explanation methods used. Secondly, through the performance's decrease, to evaluate whether the explanation fits the model's behaviour. To assess models with distinct internal logic, the validation was conducted on a more transparent model and more opaque one in both binary and multiclass situation. The results show the promising use of including the signal's derivative to introduce temporal dependency between samples in the explanations, for models with simpler internal logic.

Keywords: Machine Learning; Time Series; Heartbeat Classifier; Explainable Artificial Intelligence; Model-Agnostic Method

Resumo

As doenças cardiovasculares são, a nível mundial, a principal causa de morte e o seu tratamento e prevenção baseiam-se na interpretação do electrocardiograma. A interpretação do electrocardiograma, feita por médicos, é intrinsecamente subjectiva e, portanto, sujeita a erros. De modo a apoiar a decisão dos médicos, a inteligência artificial está a ser usada para desenvolver modelos com a capacidade de interpretar extensos conjuntos de dados e fornecer decisões precisas. No entanto, a falta de interpretabilidade da maioria dos modelos de aprendizagem automática é uma das desvantagens do recurso à mesma, principalmente em contexto clínico. Adicionalmente, a maioria dos métodos inteligência artifical explicável assumem independência entre amostras, o que implica a assunção de independência temporal ao lidar com séries temporais. A característica inerente das séries temporais não pode ser ignorada, uma vez que apresenta importância para o processo de tomada de decisão humana.

Esta dissertação baseia-se em inteligência artificial explicável para tornar inteligível a classificação de batimentos cardíacos, através da utilização de várias adaptações de métodos agnósticos do estado-da-arte. Para abordar a explicação dos classificadores de séries temporais, propõe-se uma taxonomia preliminar, e o uso da derivada como um complemento para adicionar dependência temporal entre as amostras. Os resultados foram validados para um conjunto extenso de dados públicos, por meio do índice de Jaccard em 1-D, com a comparação das subsequências extraídas de um modelo interpretável e os métodos inteligência artificial explicável utilizados, e a análise de qualidade, para avaliar se a explicação se adequa ao comportamento do modelo. De modo a avaliar modelos com lógicas internas distintas, a validação foi realizada usando, por um lado, um modelo mais transparente e, por outro, um mais opaco, tanto numa situação de classificação binária como numa situação de classificação multiclasse. Os resultados mostram o uso promissor da inclusão da derivada do sinal para introduzir dependência temporal entre as amostras nas explicações fornecidas, para modelos com lógica interna mais simples.

Palavras-chave: Aprendizagem Automática; Séries Temporais; Classificador de Batimentos Cardíacos; Inteligência Artificial Explicável; Método Agnóstico

Contents

List of Figures

List of Tables

Acronyms

k-NN	*k*-Nearest Neighbour
AAMI	Association for the Advancement of Medical Information
ADNI	Alzheimer's Disease Neuroimaging Initiative
AI	Artificial Intelligence
AIMS	Anaesthesia Information Management System
BCW	Breast Cancer Wisconsin
CAM	Class Activation Map
CBR	Case-Base Reasoning
CDSSs	Clinical Decision Support Systems
CNN	Convolutional Neural Network
CNNs	Convolutional Neural Networks
DNN	Deep Neural Network
ECG	Electrocardiogram
EHR	Electronic Health Record
FS	Free Sound dataset
GDPR	General Data Protection Regulation
Grad-CAM	Gradient Weighted - Class Activation Map
HCI	Human-Computer Interaction
HCP	Human Connectome Project
HIMSS	Healthcare Information and Management Systems Society
HP	Hepatic Patient dataset
ICU	Intensive Care Unit

ILP	Indian Liver Patient dataset
LIME	Local Interpretable Model-agnostic Explanations
LRP	Layer-wise Relevance Propagation
MM	Mammography Mass
NN	Neural Network
NNs	Neural Networks
PDP	Partial Dependence Plot
PFI	Permutation Feature Importance
PSI	Permutation Sample Importance
ReLU	Rectified Linear Unit
RF	Random Forest
SamMD	Software as Medical Device
SHAP	SHapley Additive exPlanations
UCI	University of California, Irvine
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 Context

According to the World Health Organization [1], cardiovascular diseases are responsible for 31% of worldwide deaths, each year. Being the leading cause of global death, treatment and prevention for cardiovascular diseases, rely on monitoring data and pattern evolution on patients.

Electrocardiogram (ECG) is one of the prevailing exams for triage and diagnosis as it is a non-invasive and inexpensive procedure that assesses the heart's function through its electric activity. Nowadays, the analysis of ECG waveforms is done manually by the cardiologist or technologist, a task that is prone to subjective errors and observer's variability [2]. Different sources of inconsistency while interpreting the ECG can occur when a physician reads consecutive heartbeats of the same individual or amongst doctors considering the same heartbeat [3], [4].

Artificial Intelligence (AI) has shown the ability to impact medicine at various levels: the **healthcare systems**, by improving workflow and potentially reducing medical errors; the **clinicians**, via rapid and accurate biosignal analysis; and the **patients**, by enabling them to process their data to promote health [5], [6]. More specifically, it can be advantageous to assist ECG analysis given that AI can interpret large datasets, find patterns, and make accurate predictions. Although the development of machine learning models to assist heartbeat classification is thriving, it is still in the early stage of their implementation [7].

A survey conducted by the Healthcare Information and Management Systems Society (HIMSS), in 2017, included 85 hospitals and showed that AI has its highest application in Clinical Decision Support Systems (CDSSs), and its potential use to manage data, allow early detection or produce treatment [8]. It was revealed that only 5% of the clinical

institutions use AI, and despite half of them show the intention of leveraging it daily, there is still a high prevalence of uncertainty about when to start, as highlighted in Figure 1.1. The barriers to adopt these models are mostly due to non-existing executive and physician buy-in, and lack of trust [8].

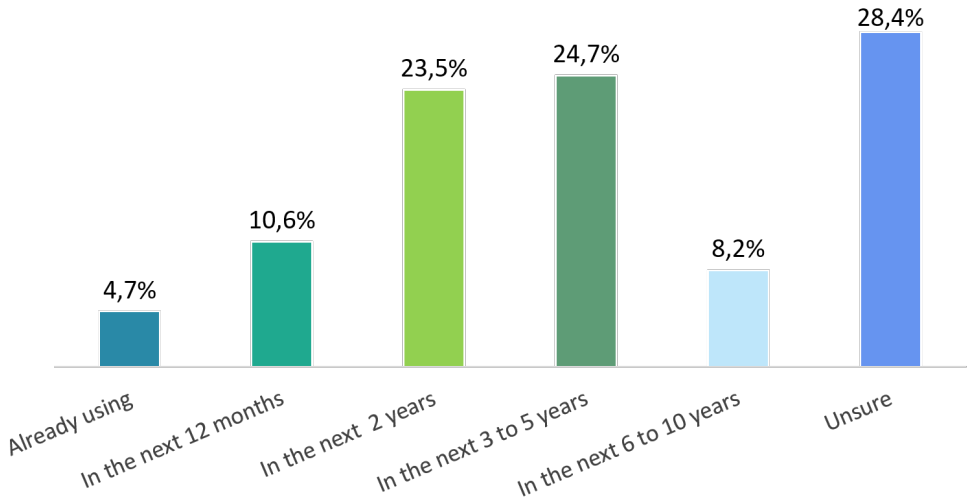


Figure 1.1: Expected leverage of artificial intelligence in the medical context. Half of the survey's participant hospitals shows the intent of deploying AI in the next five years and 5% already using it. Adapted from [8].

1.2 Motivation

Despite the benefits, the lack of interpretability of some machine learning approaches is a major drawback in their application [9]. Interpretability is a complex topic, for instance, considering a simple decision tree, it is undemanding to explain the decision by following its tree path, but when dealing with a decision tree with numerous nodes, that process is unfeasible as it becomes quite incomprehensible to humans [10]. Furthermore, the overall performance of deep learning methods has led to the proliferation of Deep Neural Network (DNN) approaches, that despite providing accurate predictions, present themselves as black-box models, as their inner logic is opaque to the users. Even though the underlying mathematical principles supporting black-box approaches are clear, their decisions are not entirely understood by experts [11]. Often, the higher the accuracy, the harder it is for a machine learning algorithm to be explained [12].

Moreover, the complexity of medical data and clinical decisions reinforces the need for trusty AI decisions. Medical data is uncertain, probabilistic, imbalanced, heterogeneous and noisy with a high number of features, that is sometimes not large enough to model complex patients, turning machine learning into a challenging task by itself. Clinical decisions are based on the analysis of diverse data and need to be represented as more than a single output [13], [14].

From the recent regulatory domain, algorithmic accountability is required by the General Data Protection Regulation (GDPR), a new European Union regulation. GDPR took effect in 2018 and opened the debate over the "right to explanation", making it necessary for AI systems to have an explanatory description behind their decision [15]. To solve the existing gap between the resulting research outcomes and the society's expectation, the European Commission addressed a set of principles to guide the development of trustworthy AI, focusing on robustness and explainability. The correct research stands on transparency, encompassing interpretability and explainability, reliability, relating to consistency, and data protection in models. Some of the paths for the implementation of responsible AI include the evaluation of its impact in society, the categorisation of methods to assess its robustness, and the promotion of transparency when developing models, emphasising the need for self explainability to protect users' rights [16].

Explanations have an important role, mostly in fields where accountability is necessary and a mistake could be fatal, as it happens in the medical domain [17]. Wrong CDSSs' recommendations can mislead doctors into an incorrect decision, and hence high-stake decisions can not be disproved of explanations. Self-explaining AI is useful to identify and correct errors in the algorithm, allow a careful approach of these models and accept their recommendations, and even understand underlying physical phenomenons [18].

With the change in the AI's paradigm, it is imperative to bring explanations into the classical pipeline of machine learning, to satisfy the growing demand for models that provide not only accurate diagnoses but also justifications for their behaviours. Thus highlighting the need for a unified notion of explanation and of assessment metrics to support the incoming Explainable Artificial Intelligence (XAI) studies [19]. Scientific progress in the area of XAI is capable of improving the healthcare system, not only in the areas identified above but also in the context of regulatory approval of clinical decision support systems, since the explanations can assist regulators in evaluating if the product is complying with the regulatory requirements [20].

Past studies in the context of clinical XAI have primarily focused on computer vision tasks and natural language processing, but less often on time series. Visual explanations provide information for making inferences about the behaviour, causality, providing knowledge about the function of the system [21]. To understand and analyse predictions of different models, visual explanations can be considered an initial approach to explain time series, as they show correlations between features and outline the importance of the instance's morphology [22].

1.3 Literature Review

This section includes the literature review of the latest methods to explain the behaviour of classifiers in a clinical context. Firstly, the historical background of XAI is highlighted and finally, the latest studies upon several types of clinical data are described.

1.3.1 Historical Background

Despite the recent outburst of XAI, there is a continuous history of work on the explanation's psychology [23]. The introduction of explainability in autonomous decisions emerged twenty years after the first AI historical publications of Turing [24], Minsky, Edmonds [25] and Samuel [26]. Historically, self-explaining systems can be divided into three categories: the first generation, that provided reasons for such output, the second generation, or tutoring systems, returned an additional reaction strategy; the third generation, able to explain more complex machine learning models. Both the first and second generations were named expert systems, comprised of rule-based AI approaches constructed with logical if-then rules from expert knowledge.

The first explanation strategies were developed during the 1970s and consisted of presenting the traceback of the decision by a series of boolean conditions and symbolic workflows. That is the case of MYCIN, an inference engine designed to identify bacteria that causes severe infections and issuing recommendations for antibiotics. Explanations were issue by translating the list of rules to reach the specific prediction, into a human-based speech [27]–[29]. To overcome the inadequacies of the first generation, tutoring systems were introduced in the mid-1980s, which aimed to explain not only the *why?* but also the *what to do next?* [30], [31]. For instance, a tutoring system called GUIDON, a guide for medical students, could provide constructed arguments with multiple explanations [32]. This system rested on the MYCIN knowledge base, combined with an interpreter for applying rules and with user interaction [32]. During the 1990s to the 2010s, the area of expert systems recessed, as the research of XAI encountered obstacles and suffered from a general scepticism, entering a period known as *explainability winter* [33].

1.3.2 Recent Outlook

From the 2010s onwards, technological development such as the development of graphic processing units motivated the application of heavier computational techniques [34]. For instance, allowing the implementation of more complex AI models, such as deep learning, more difficult to understand. Hence, the need for explanations emerged again, giving place to the third generation of XAI. Since then, several researchers attempted to define explanations methods and proposing high-level taxonomies for XAI [35]–[38]. Some of the recent activity in XAI has been prompted by the GDPR, as it has claimed the *right to explanation* and thus, required additional information about AI systems' decisions [39].

Recent papers sustained that the future of XAI stands on a co-creation including the developer and the end-user, in order to provide an explanation that is consistent with its specific use, dividing the field into two main groups: algorithmic-centric, suggesting interpretability methods that discard human subject tests; and user-centric, propelling the Human-Computer Interaction (HCI) community to develop different explanation methods, such as a question-driven explanation [40]–[42]. Furthermore, the need for precise

and consensual metrics to assess an explanation was approached by Doshi-Velez et al. [19], with a proposed *taxonomy of interpretability evaluation*. It is suggested that three levels of evaluation exist. Two of them requiring human competences, the application-grounded, the most expensive as it requires expert domain subjects, and human-grounded, relying on simpler tasks. The least specific, functionally-grounded evaluation, exempts human tasks and stands on a formal definition of interpretability as a proxy for explanation quality.

1.3.3 Explainable Artificial Intelligence in Medicine

Most of the current research in XAI has been focusing on the development of methods for computer vision and natural language processing tasks. There are relatively few works on literature that systematically address time series data and even less prevalent in clinical time series. In medicine, attempts to explain CDSSs have been explored, either focusing on image, tabular, and text data.

Table 1.1 summarises the most recent works addressing the topic of XAI in clinical data. Despite focusing on a more comprehensive scope for time series, we also provide examples of previous work in the context of image and text data. We categorised each study according to the specificity and type of data. The specificity criterion refers to the classification of XAI methods as model-specific if their development and application are dependent on a model's internals, or as model-agnostic if they can be applied to any classifier. The XAI method and dataset are also presented.

1.3.3.1 Image Data

In the context of image classification, explanation methods often return the relevance of each pixel or superpixel for the classification. XAI studies based on clinical data explained deep learning approaches through Class Activation Map (CAM) [43], [44], Layer-wise Relevance Propagation (LRP) [45] and saliency map techniques [46], to highlight the image regions that most contribute to the classification. LRP consists of a deep Taylor decomposition function applied along with hidden layers of Neural Network (NN)s, which associates each decomposed region with a coefficient [31]. CAM is obtained by the dot product of the extracted weights from the final convolutional layer and the feature map. The map is upsampled and superimposed on the input image to show the regions that the Convolutional Neural Network (CNN) considers most important [47].

Explaining image classifiers have also shown to be useful to expose hidden misclassifications, as Zech et al. [44] presented, in which, the activation maps, showed that NN could rely on subtle differences of the image processing, or even compression information and ignore real pathology characteristics. Lapuschkin et al. [48] referred to the topic of overfitted models as the *Clever Hans effect*, a psychological phenomenon of learning characterised by the influence of external information. Through LRP, the accurate classifications demonstrated high relevance in the source tag to identify the type of object,

instead of their specific characteristics.

1.3.3.2 Tabular Data

Regarding tabular data classifications, explanations were explored by different model-agnostic methods. For instance, by globally replacing the complex model by an interpretable model to explain Intensive Care Unit (ICU) prognostic prediction [49], [50], by using LRP for acute critical illness and stroke outcome prediction [51], [52], or by Case-Base Reasoning (CBR) to explain breast cancer prediction [53]. CBR consists of justifying a model's decision with the most similar instance in the training dataset.

More precisely, Che et al. [49], explained the prediction of mortality after 60 days and Rafi et al. [50], explained the prediction of 30-day ICU readmission, following similar approaches. The classifier used was a combination of recurrent NN that dealt with temporal data and DNN to deal with the categorical features. To explain this model, an intrinsically interpretable model was used to mimic its behaviour. The interpretable model applied was the gradient boosting tree, decision trees with the ability to optimise a cost function by iteratively choosing the direction with a negative gradient, whereupon a set of interpretations was performed such as, feature importance, by inspecting the trees and with the observation of partial dependence plots, that give insights about how a certain feature impact the outcome.

Local Interpretable Model-agnostic Explanations (LIME) has also shown its application in the medical domain. LIME is a model-agnostic and perturbation-based method that attempts to understand the model by perturbing the input's vicinity and measuring the impact of how the predictions change. LIME produces local explanations for single predictions through which the variation produced on the classification represents a relevance score for each sample. The success and robustness of LIME's explanation are influenced by the sample's vicinity, dependent on the kernel's width. A deterministic adaptation of LIME, based on hierarchical clustering for the perturbations, was developed by Zafar and Khan [54] to allow a more consistent and stable behaviour for computer-aided diagnosis.

1.3.3.3 Text Data

In the domain of natural language processing, most of the methods applied are model-specific methods, such as LRP and attention mechanisms, which have shown promising results [55]. Mullenbach et al. [56] explained medical codes prediction from the clinical text written after the patient's encounter with the physician. The medical codes translate information about the diagnosis and advised treatment. Explanations are obtained through an attention mechanism, that identifies which part is the NN focusing on and which features influence its choice. Attention mechanisms combine network activations in the latent space of the sequence into a set of learned attention weights. Explanations

for text data classifiers are the closest to time series, as their instances are expressed as a vector of features, upon which the order of each feature holds importance.

1.3.3.4 Temporal Data

Regarding deep learning approaches on temporal data model-specific methods, such as CBR techniques [57], and attention mechanisms [58], [59] have been used. Gee et al. [57] proposed the use of prototypes to expose representative morphologies for classifiers using different types of clinical data, allowing them to understand ECG morphology while classifying bradycardia events, respiratory waveforms when classifying apnea and audio waveforms for spoken digits classification. Prototypes are encoded in the deep classifier, providing intrinsic explanations for their behaviour [60]. Lin et al [59] used attention-based temporal CNNs to explain myotonic dystrophy diagnosis, a progressive neuromuscular disease characterised by the delayed muscle relaxation after its contraction. The provided explanations consisted of the most relevant segment from the handgrip time series from a patient to reach the model's decision. LRP has also been used to explain DNN classification in text [61] and temporal data [62]. In the former case, to get the relevance of the words in newspaper classifier, and the latter, to explain an individual classifier by providing the importance of subsequences of the gait pattern analysis. Horst et al. [62] showed a possible interpretation of such black-box models and uncovered features that express the uniqueness of individual gait patterns.

Model-agnostic approaches were explored as well. On one hand, Slunderberg et al. [63] explained the prediction of hypoxaemia during surgery using features extracted from time series, through the use of SHapley Additive exPlanations (SHAP). SHAP is based on game theory to calculate the Shapley Values as relevance's scores. Time series features were extracted from data with uneven sample rates previously combined: data from the Electronic Health Record (EHR) and Anaesthesia Information Management System (AIMS). Data were then translated into a vector, in which static information was repeated after its measurement, and both time series and drug administration data were represented by an exponential decay with different decay rates. On the other hand, Mujkanovic [64] and Guillemé et al. [65] both introduced adaptations of SHAP and LIME into explaining time series classifiers using the raw signal. A disadvantage of perturbation-based methods is that they omit temporal dependencies as they assume the independency between samples, and thus produce explanations only partially verifiable on time series data.

To assess the quality of the explanations using a functionally-grounded, there is only limited work. In the work by Guillemé et al. [65] they proposed a method to measure fidelity by comparing an explanation from the chosen method against an interpretable classifier, using Shapelets. Shapelets are the most discriminative subsequences from a time series, to identify classes [66]. On the other hand, Mujkanovic [64] compared the explanations from the adapted SHAP from different time series classifiers, through the

median correlation. Furthermore, Schlegel et al. [67] 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 method are replaced. The higher the drop in the model's performance, the more reliable and representative the explanation is. Several replacement functions were considered, such as replacing by zero, the mean value, or swap, which requires swapping the order of the samples amongst that sequence. The last replacement allowed 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.

1.3.4 Summary

XAI methods for machine learning are not recent and have always had an important role in AI, as reviewed in section 1.3.1. The recent researches on clinical data are usually divided into two stages: first, the development of the prediction model with high accuracy, and secondly, the development and application of explanation methods.

The related work highlights the latest effort in including XAI methods into clinical time series classifiers but also exposes the lack of model-agnostic applications. Furthermore, the existing model-agnostic methods are still far from optimal as they assume temporal independence. Moreover, it becomes important to define a system for categorising time series explanations and a validation protocol, to enable the support of our study. Accordingly, a rigorous evaluation of the chosen dataset is relevant to create explanation methods which are able, not only to provide explanations consistent with the complex model but also to produce meaningful justifications for domain experts.

Table 1.1: Description of recent XAI studies on clinical data.

Study	Specificty	Method	Type of Data	Data
Thomas et al. (2019) [45]	Model-specific	Adaptation of LRP[1]	Image data	fMRI dataset of HCP[2]
Yang et al. (2019) [43]	Model-agnostic and Model-specific	Sensitivity analysis and CAM[3]	Image data	Brain MRI scans from ADNI[4]
Zafar and Khan (2019) [54]	Model-agnostic	Deterministic adaptation of LIME[6]	Tabular data	UCI[7] repository, (1) BCW[8], (2) ILP[9] dataset, (3) HP[10] dataset
Mullenbach et al. (2018) [56]	Model-specific	Attention mechanism	Text data	Discharge summaries from MIMIC-III
Lin et al. (2019) [59]	Model-specific	Attention mechanism	Time series data	Hand-held dynamometer from handgrip strength
Gee et al. (2019) [57]	Model-specific	Learned prototypes	Time series data	Neonatal ICU[13] dataset, ECG[14] and Respiration waveforms ; FS[15] dataset
Horst et al. (2019) [62]	Model-specific	LRP[1]	Time series data	gait data, lower-body joint angles and ground reaction forces
Slundberg et al. (2018) [68]	Model-agnostic	SHAP[11]	Time series data	AIMS[12] and EHR[5] from hospitals
Guillemé et al. (2019) [65]	Model-agnostic	LIME[6] and SHAP[11]	Time series data	UCR time series classification archive
Mujkanovic et al. (2019) [64]	Model-agnostic	SHAP[11]	Time series data	UCR time series classification archive
Schlegel et al. (2019) [67]	Model-agnostic and Model-specific	LIME[6] and SHAP[11]; LRP[1], saliency and DeepLIFT	Time-series data	UCR time series classification archive and MIT-BIH arrythmia database

[1]Layer-Wise Propagation [2]Human Conectome Project [3]Class Activation Mapping [4]Alzheimer's Disease Neuroimaging Initiative
[5]Electronic Health Record [6]Local Interpretable Model-agnostic Explanations [7]University of California, Irvine
[8]Case-Base Breast Cancer Wisconsin [9]Indian Patient Liver [10]Hepatic Patient [11]SHapley Additive exPlanations
[12]Anaesthesia Information Management System [13]Intensive Care Unit [14] Electrocardiogram [15] Free Sound

1.4 Objectives

This dissertation presents an exploratory study on explaining ECG classifiers, giving insights about XAI methods to justify the behaviour of these models. To address that, we propose an initial taxonomy towards standardisation of XAI methods' development for time series, followed by an adaptation of the state-of-the-art methods optimised for time series data. Finally, we validate our modifications with real clinical ECG data.

In this context, one can identify several research directions towards the successful application of XAI methods in time series. The main research question is: **How to explain a time series classifier?** Several other resulting questions arise, such as: What are the most important requirements when explaining a time series classifier? Would it be feasible to adapt state-of-art XAI methods to time series? Are the used methods sensitive to the order of the event's occurrence? We argue that in the context of explaining time series classifiers, one must take into account the temporal dependency between samples or multiple time series. Therefore, the XAI methods must allow exposing temporal dependency, resembling the human decision process, which sometimes is grounded in interpreting time series by not only evaluating the amplitude but also, in the temporal order of the events.

Accordingly, the more specific objectives to answer the main question are to (1) define the requirements for explaining time series classifiers, (2) apply a set of modified explanation methods across different machine learning models with an incremental level of complexity, (3) create a validation protocol, to evaluate the explanation's quality and finally, (4) validate such approaches on an extent ECG dataset.

By justifying the machine learning model's behaviour through explanations, it is possible to increase confidence in its recommendation, raise awareness for its bias, and uncover hidden physiological phenomenons. Ultimately, to facilitate their deployment in the medical domain.

1.5 Structure

This document is composed of five chapters as represented in Figure 1.2.

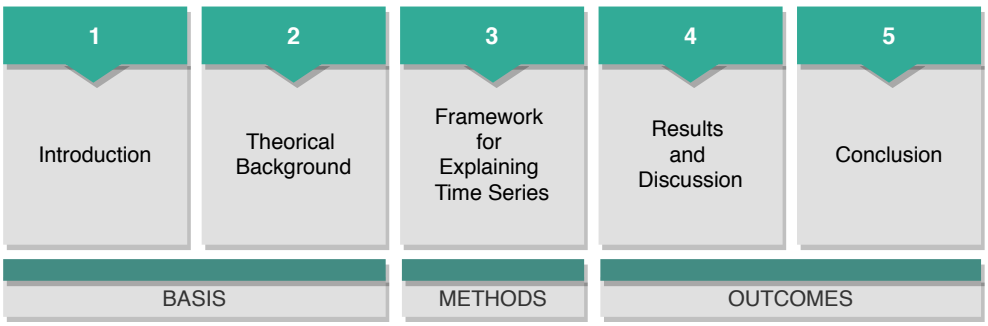


Figure 1.2: Dissertation structure overview. Consists of five chapters, the introduction, theoretical concepts, framework, results and conclusion.

The covered chapter, Chapter 1, presents the context and motivation behind this dissertation, and its main objectives. Chapter 2, addresses the necessary theoretical concepts to deal with XAI on ECG classifiers. Chapters 1 and 2 compose the basis of this dissertation. Chapter 3 includes the details about the followed framework to approach time series explanations, namely the used methods and the necessary adaptations, a proposed taxonomy, and a description of the validation tools. Lastly, the outcomes are presented, including the results and discussion, addressed in Chapter 4, and the conclusion along with some future work suggestions, in Chapter 5.

Chapter 2

Theoretical Background

This chapter presents a broad overview of the theoretical concepts in the context of this dissertation. First, the underlying physiological principles about the electrocardiogram are approached, along with the description of the heart's normal functioning. Secondly, the theoretical foundations of machine learning and explainable artificial intelligence are presented, focusing particularly on model-agnostic methods, the ones approached in this dissertation.

2.1 Electrocardiogram

The ECG is a standard non-invasive exam that monitors the heartbeat's rhythm and gives insights about the size and position of the heart chambers. ECG's waveforms translate physiological representations of the heart, in terms of structural and functional information about the heart. They are used for the diagnosis and prevention of cardiovascular diseases. Moreover, the waveforms are a graphical representation of the heart's electrical activity measured over time, obtained by electrodes placed on the skin. The cardiac signal is generated by the depolarisation and repolarisation of the heart's muscle tissue, which allows it to contract and pump the blood to the various body regions. The sequential occurrence of the heart's contraction composes the cardiac cycle.

The heart has four chambers, two atria, and two ventricles, composed of different types of cells, i.e., the contractile muscle cells, named cardiomyocytes, and the conduction cells. The first cells are responsible for providing contractility, and the latter form the conduction system and are responsible for the generation and conduction of the action potential. During each heartbeat, the potential, initiated by the pacemaker cells, is propagated across the conduction system. Pacemaker cells are self-sustaining rhythmic cells, that allow the heart to maintain a paced rhythm. The stimulus they produce, the

action potential, depolarises the contractile cells from the atria to the ventricle, enabling atrial and ventricular polarisation and depolarisation [69]. Thus, providing the ECG its specific representation.

An ECG is the sum of all the electrical activity from the stimulated areas of the heart. Normal cardiac cycles usually consist of several segments, and five waves, P, Q, R, S, and T, as presented in Figure 2.1 by the cardiac cycle.

Initially, the P wave is originated by the contraction of the atria myocardium, followed by the depolarisation of the ventricles, the QRS complex. Lastly, it comes the T wave, caused by the repolarisation of the ventricle during ejection [70]. Additionally, although not observed very often, there is the U wave after the T wave, not represented in the illustration.

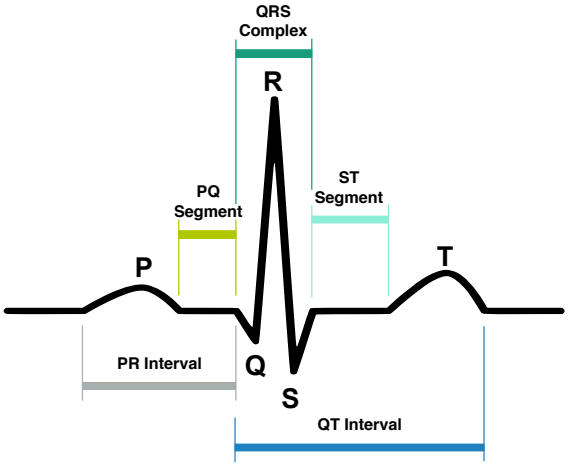


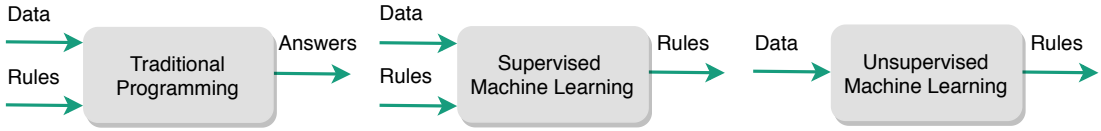
Figure 2.1: Illustration of a normal heartbeat and its segments. The P wave is followed by the QRS complex, and T wave.

2.2 Machine Learning

Machine learning is the application of AI that enables self-learning and improvement from data, through algorithms. Machine learning models learn from statistical patterns in a multidimensional space, such as patient analysis or a list of symptoms, and make predictions based on that [71]. Instead of needing a large list of rules as the traditional methods require (Figure 2.2a), machine learning is able to acquire knowledge and improve its own code from the provided data [72].

Usually, machine learning approaches are categorised into three main groups: supervised (Figure 2.2b), unsupervised (Figure 2.2c), and semi-supervised, based on whether the training data is labelled or not.

In this dissertation, we will focus solely on supervised methods, whereas the model learns a function from the provided pair of inputs and outputs, and is able to generalise into classifying new inputs [74].



(a) (b) (c)

Figure 2.2: Traditional programming compared to supervised and unsupervised machine learning programming. (a) Traditional programming (b) Supervised Learning, and (c) Unsupervised Learning. Adapted from [73]

The dataset D, and labels y, are both matrices that represent the inputs, X, and the respective targets from which the classifier learns [72]. The dataset is split into the training and test set to enable generalising the learned model without compromising its evaluation. These models generate an approximation function, $\hat{f}$, from the training data. This hypothesis of the real function, f, yields a prediction, $\hat{y}$, based on the input variables. Further adjustments are then made to minimise the error between the prediction and the expected class, granting the classifier's improvement [75]. The model's prediction is described by:

$$\hat{y} = \hat{f}(X) \tag{2.1}$$

Evaluating the model's performance stands on assessing its behaviour on data that was not used on the training stage, from the test set.

Regarding the criterion of interpretability, machine learning models can be classified as interpretable or non-interpretable. Interpretable models can be understood by humans, such as linear and logic regression, decision trees, Naive Bayes, and k-nearest neighbour [72]. Contrarily, the non-interpretable models, also named black-box models, lack transparency as their inner logic is not easily perceived by inspecting the internal parameters, such as DNN [76].

It is often true that there is a commutation between accuracy and interpretability, in the way that models with higher accuracy tend to be more complex and thus harder to explain. Higher model complexity accounts for higher flexibility, such as the computation of non-linear relations between features, which is less interpretable, as it makes the cause and effect harder to understand [37]. On both extremes are classification rules, the most interpretable and least accurate methods, and neural networks, the most accurate but least interpretable.

2.2.1 Interpretable Models

Interpretable models are a subset of algorithms in which their decisions can be comprehended by humans. In other words, model interpretability requires that each step towards the outcome is traceable.

Interpretable models are relevant to uncover causal structure in data [10]. A casual

relationship between variables is often misled with correlation but these are two different concepts. If two features are correlated, it does not imply that one is the cause of the other. A high correlation between features only highlights their similar growth or decrease over time. Unlike correlation, which can have no reliable information about the input relationships, causality is a property that provides valuable knowledge about a system [77].

2.2.1.1 Linear Regression

Linear regression is an algorithm that performs a regression task between the input variables and the output, which means that the target is described as a linear combination of all the features. Linearity makes it possible to easily understand and interpret the system, but it can create unreliable representations of the real phenomenon.

The learned relationships can be described as:

$$\hat{y} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon \tag{2.2}$$

where x represents the input variables or features, β are the weights or coefficients, $\hat{y}$ the output or prediction, and ϵ is the difference between the prediction and the real target. The optimal weights can be estimated using different approaches, but they usually tend to search for values that minimise the error. The estimated weights provide relevant information and allow us to interpret the trained model since different weights represent the influence that a specific feature has in predicting the outcome [72].

An important measurement for interpreting linear models is the R^2 as it gives information about how much of the total variance of the target can be explained by the model. Low values of R^2 can provide misleading information because such model is not capable to explain much of the variance. Any interpretation of the weights would not be a reliable representation [72]. However, each non-linearity and interaction needs to be handcrafted, often leading to models with low performance, as the relations between features are oversimplified.

2.2.1.2 Decision Tree

Tree-based models resemble the tree shape in which the branches represent decisions or reactions. According to certain cutoff values in the features, the data is recursively split into smaller subsets according to the tests set in the branches. While following through different branches and nodes, different subsets are created, until it reaches the final node, called terminal node or leaf node, and a decision is made. To reach a predicted outcome, the intermediate nodes rely on the statistic values of the training data by using its average [78]. An example of such models is portrayed in Figure 2.3. Despite the variety of the algorithms that create decision trees, they mostly depend on purity criterion, such as Gini criteria, to better choose how to divide the data. This is relevant because reducing the resultant impurity implies a lower entropy which will increase gain function [72].

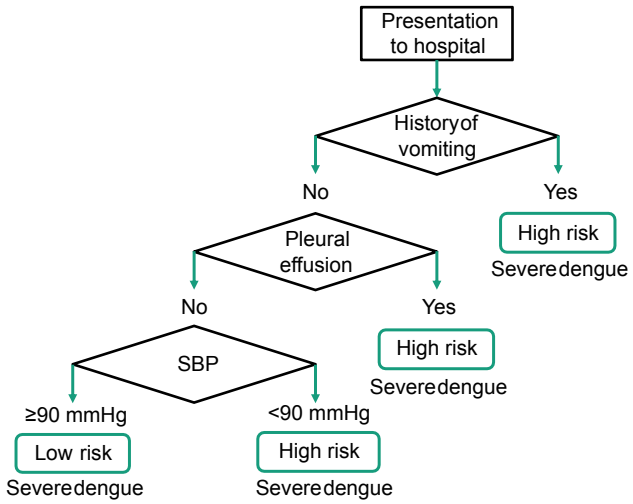


Figure 2.3: Schematic example of a decision tree model. In this example, the classes are the presence or absence of Dengue. Adapted from [79].

The relationships can be described as:

$$\hat{y} = \sum_{m=1}^{M} c_m I\{X \in R_m\} \tag{2.3}$$

where c_m is the average of the training data in that node, and I corresponds to an activation function that returns 1 if the X is a part of the subset R_m or 0 if not. The predicted output corresponds to the average value of the training data of the terminal node.

Interpretation for a single prediction is about following the path of the input X, from the first node until the leaf node. Feature importance can also be calculated in this case, returning the relevance of each model's feature, by calculating in each split how much the entropy has decreased. The more the entropy drops, the more important the feature [72].

Decision trees are optimal to learn non-linear relations between features and for capturing interactions. However, they are prone to overfit if parameters as depth and number of nodes are not properly controlled, and are unstable as a small change in the dataset can change the tree structure [72].

2.2.1.3 RuleFit

The rule fit algorithm is a sparse linear method that has the same working principle as the linear regression with the advantage of being able to consider interactions between features, as described in Figure 2.4. RuleFit generates decision trees with rules, and through those trees learns a linear model. Rules can be understood as functions, that will ultimately be turned into new features. Rulefit often generates a large number of rules from the training set. Hence, a sparsity method is applied, to reduce the number of features. A sparse linear method aims to reduce dimensions, by setting in Equation 2.2, several β values to zero [80].

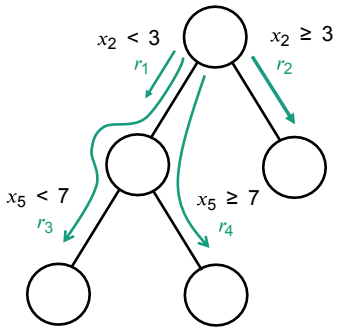


Figure 2.4: Schematic example of a RuleFit model. Rules are interpreted features. Adapted from [72].

For its interpretation the RuleFit returns the weights and respective importance, providing an understanding of how the features change the prediction.

On one hand, RuleFit solves the problem of linear models by including feature's interactions, and on the other, improves interpretability because it is a sparse method. However, even applying methods to reduce features, the features created can be too many and generate nonessential rules [72]. Often interpretable models face this obstacle, as they easily generate numerous explanations in the form of rules, traces or important features. Notwithstanding, the number can be so large that it makes interpretation impossible as humans are not able to absorb and fully understand the interpretable method.

2.2.1.4 *k*-Nearest Neighbour

k-Nearest Neighbour (*k*-NN) is an instance-based classifier, which means that it makes decisions not based on learning distributions, but based on the training samples stored in the memory. In other words, *k*-NN predicts classes based on the classification of the $k \in \mathbb{N}$ nearest instances from already known data [72]. To measure the similarity of two instances and select neighbours, the *k*-NN uses distance metrics, such as the Minkowski and Cosine distances. Additional distance metrics are presented in Table 2.1. Manhattan and Euclidean distances are derived from the Minkowski distance with $p = 1$ and $p = 2$, respectively.

Time and space consumption for obtaining the predictions is often high, as it is necessary that all the distances between the predicted instance and the training set are calculated.

The choice of the number of neighbours must be suited to the type of data. In the case of a small number of neighbours, the noise will have a higher influence on the result, and a large number of neighbours make it computationally expensive. The number of neighbours is related to variance and bias. A small number of neighbours are a most flexible fit, hence having a lower bias but higher variance and a large number of neighbours produces smoother decision boundaries which means lower variance but higher bias [81].

Neighbours can be used in different ways to determine the instance class. For instance,

Table 2.1: Commonly used distance metrics, to calculate the distance between two instances X and Y.

Distance	**Metric**
Minkowski$_{orderp}$	$D(X,Y) = (\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p)^{\frac{1}{p}}$
Euclidean	$D(X,Y) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$
Manhattan	$D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
Cosine	$D(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$

the predicted value can be based on the majority, in a classification task, or the mean value of the neighbour's classes, for a regression task. Additionally, the instances can be weighted to assign more weight to the nearest neighbours in deciding the class [82].

k-NN can have its predictions explained by retrieving the nearest instances. However, to interpret an instance with hundreds of samples, interpretability becomes debatable, as it would be necessary to explain every single sample. In this case, it is necessary to reduce the number of considered samples in the explanation to convert them into a human intelligible format [72].

2.2.1.5 Shapelet-based Classifier

Shapelet-based classifiers are based on the Shapelet representation of the dataset to discriminate classes. Shapelets are defined as the subsequences that can maximally describe a class [66]. This technique was developed to overcome some of the challenges of time series classification, namely high computing time and occupied space of instance-based classifiers. Shapelets can determine similarity based on smaller discriminative shapes, instead of using the entire time series length.

To find the Shapelets, the brute-force search is based on an exhaustive search of all the time series' candidates and thus, suffers from high runtime complexity. A series of speed-up techniques were proposed, based on the early abandon of distance computations and entropy pruning of the information gain metric [66]. The initial approach to leverage the advantage of using Shapelets for time series classification was based on tree-based classifiers [66]. For instance, in a binary classification case, the Shapelets are compared with all the instance subsequences and if the distance between them is smaller than a certain splitting value, it is classified as 0, if not is classified as 1.

The traditional method for obtaining the Shapelets is computationally expensive and other versions have been proposed. Learning Shapelets is an algorithm presented by Grabocka et al. [83] for the Shapelet discovery, that instead of searching the subsequence that allows better classification performance, it learns a Shapelet from a dataset through a stochastic descent gradient. This process learns subsequences that can linearly separate

the distances from the dataset by their classes. The subsequences are learned randomly by guessing the Shapelet and are iteratively optimised to minimise the classification loss function. Minimum distances become the new predictors in the transformed Shapelets space and a linear learning model can predict approximate target values, as shown in Figure 2.5.

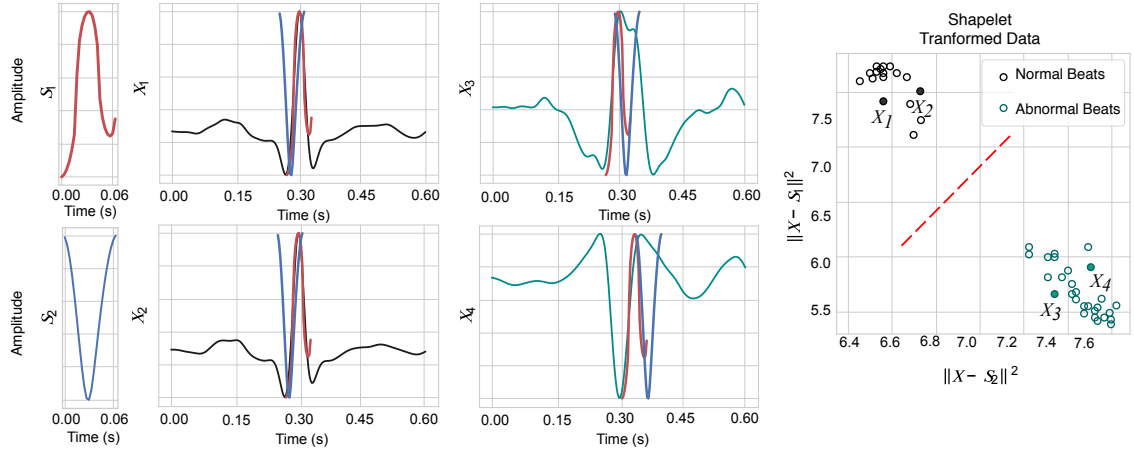


Figure 2.5: Learning Shapelets example, using two Shapelets S_1 and S_2, to perform a binary task of classifying normal and abnormal heartbeats. Adapted from [83].

It is possible to retrieve the interpretability of a Shapelet-based classifier, through the minimum distances or by visualising the extracted Shapelets, as they represent the most representative subsequences of a specific class.

2.2.2 Non-Intepretable Models

The inner logic of non-interpretable models is far more complex as it becomes incomprehensible to trace the input towards its output. These models can compute the existing non-linear interactions between inputs, which can be apprehended by deep learning models and random boosting trees [84].

2.2.2.1 Convolutional Neural Network

Neural Networks (NNs), are characterised as chains of single artificial neurons whose output is computed by a non-linear function of the summed inputs and weights. This output becomes the input for the next connection and so forth. Inspired by biological neural networks, these models are composed of multilayered artificial neurons. NNs are trained by a sequential adjustment of the weights from the output to the input, named backpropagation, to minimise the difference between the predicted and real outcome [85]. A DNN is characterised by having many hidden layers, making the output's explanation a very complex task.

A specific type of DNN is CNN, in which the hidden layers are composed of interleaved convolutional and pooling layers, in varying numbers, to generate features of the

raw data, as exemplified in Figure 2.6. A CNN consists of an input and output layer, and several hidden layers, in order to generate automatically deep features of the input [86]. The convolutional layers perform a mathematical operation on the input coming from the previous layer trough kernel filters, generating feature maps. On image data, these filters exhibit two dimensions, width, and height, and on temporal data, these filters only have one dimension, which is time [87].

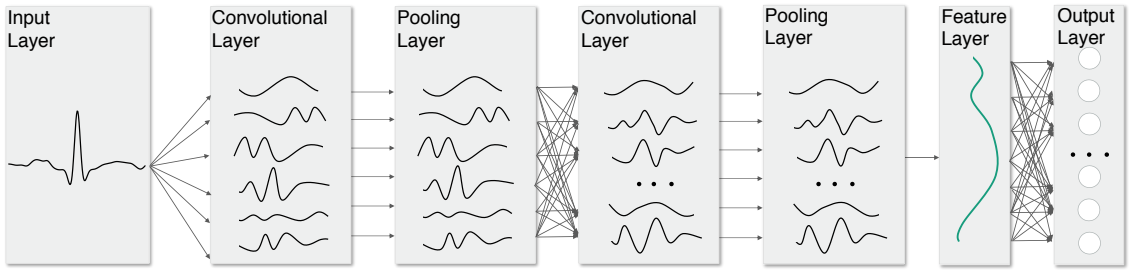


Figure 2.6: Example of a CNN with two convolutional and pooling layers, followed by the fully connected and the output layer, applied to an ECG. Adapted from [87].

The output at a convolutional layer, the feature map, is replaced by a summary statistic of the nearby outputs in the pooling layer. In other words, the pooling layer reduces the convolved features, by extracting the most relevant ones. Pooling operations can be of several types, such as the average pooling and the maximum pooling. Consequently, the original input is represented by a series of feature maps.

The feature layer is a fully connected layer, which connects all the feature maps to generate a new transformed instance and feeds this data into the output layer. Finally, the output layer, has n neurons, corresponding to the possible classes. The classification is usually done based on the maximum output, from the output layer [87].

2.2.3 Performance Metrics

The performance evaluation takes a relevant role in the development of machine learning models. Different metrics allow assessing distinct information, such as how the model behaves when predicting positively, or how many of the real positives the model identifies.

The confusion matrix is a specific matrix for the visualisation of the model's performance, which presents the true labels against the ones predicted by the model. As illustrated in Figure 2.7, the confusion matrix enables the comparison between correct predictions and the incorrect ones, according to the different classes.

Several metrics can be derived from this representation, such as the accuracy, precision, recall, and F_1 score.

1. **Accuracy:** Accuracy is simply given by the fraction of the correct predictions and total predictions. The accuracy score can be quite uninformative when dealing with unbalanced datasets given that it is possible to have high scores even if the model

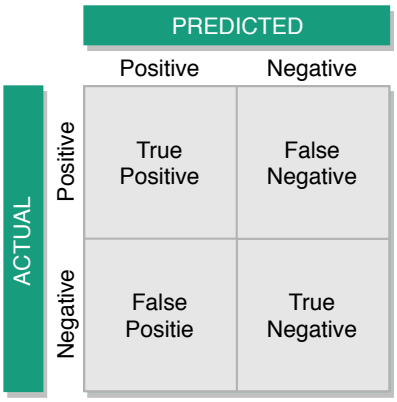


Figure 2.7: The confusion matrix, exhibits the predicted class in the rows an the real class in the columns.

is not performing well in the smallest class [88].

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{2.4}$$

Where T_P refers to true positives, T_N to the true negatives, and F_P and F_N denotes for false positives and false negative, respectively.

2. **Precision:** Precision assesses the positive predictions. More precisely, this metric is calculated by the ratio between the real positive cases and predicted positives, including both correct and false predictions.

   Precision is defined by Equation 2.5:

$$Precision = \frac{T_P}{T_P + F_P} \tag{2.5}$$

3. **Recall:** The recall score represents the predicted positives in comparison to the true labels. It is defined as the fraction between the positively predicted classes and the real positive cases existing. Recall can be determined by:

$$Recall = \frac{T_P}{T_P + F_N} \tag{2.6}$$

4. **F_1:** The F_1 score is the harmonic mean between precision and recall and balances the information from those metrics.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.7}$$

   F_1 score allows evaluating the performance on unbalanced datasets due to the fact that it compares both the recall and precision. A model with high precision but low recall score returns very few positive results, but most of its predicted labels are correct when compared to the ground-truth, and a model with high recall but low precision score returns many positive results, but most of its predicted labels

are incorrect when compared to the ground truth. A model with high precision and high recall is ideal as it means many predictions are returned with all results labelled correctly.

Another metric that can be used to assess the model's performance is the Jaccard index. The Jaccard index or Jaccard similarity coefficient is a score that evaluates the similarity or diversity between two datasets. This metric measures resemblance among two finite sets, A and B, and is represented by the size of their intersection divided by the size of their union.

The Jaccard index is given by:

$$J(A,B) = \frac{\#A \cap B}{\#A \cup B} \tag{2.8}$$

where # is the cardinality or size of the set.

2.3 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) is the field of AI that focuses on the development of transparent and trustful machine learning models. Explanations are important when dealing with automatic decision algorithms in the various domains, for different reasons: **to allow a possible justification**, since accountability must be present to foster a responsible AI culture; **to control and improve**, so it can evidence the algorithm's strengths or weaknesses; **to discover**, to reveal complex mechanisms by showing unknown patterns in the dataset [36].

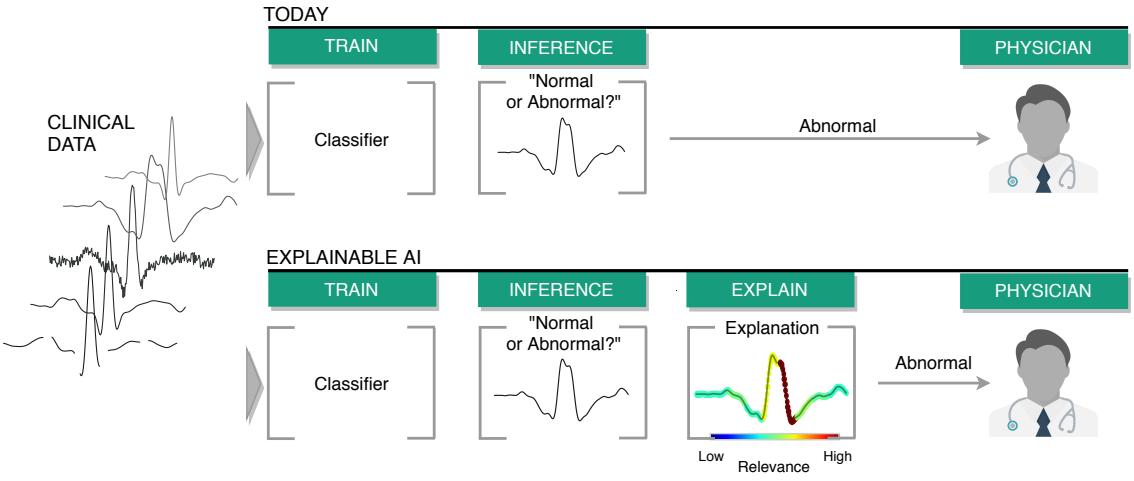


Figure 2.8: Comparison between the traditional classification pipeline (on top) and the XAI pipeline (bottom), applied to the medical domain. The pipeline from XAI introduces explainable models and interfaces. In this case, the inference is explained by measuring the relevance of each sample of the time series for the outcome.

XAI refers to methods through which the resulting prediction is explained in a human intelligible format. It aims to provide effective explanatory techniques [89]. Figure 2.8

illustrates the need for self-explaining machine learning methods and enlightens XAI's pipeline, which adds an interface to explain the predictive model. This ideology results in trustable systems enabling the user to understand why and how a prediction was made, instead of the regular machine learning pipeline, that is more focused on providing outputs [89].

2.3.1 Interpretability and Explainability

Despite the efforts to unify a formal description, there are still no rigorous definitions in XAI, and ambiguity amongst the terminology remains [19], [48], [90]. For instance, interpretability and explainability are often used interchangeably, but many authors describe them as two different concepts [35], [36]. The slight difference between these concepts is denoted:

> *An interpretation is the mapping of an abstract concept, (e.g. a predicted class) into a domain that the human can make sense of. An explanation is the collection of features of the interpretable domain, that have contributed to a given example to produce a decision (e.g. classification or regression) [31].*

Interpretability is about the extent of being able to fully understand the cause and effect in a system. Interpretable domains are, for instance, images or texts. Explainability relates to uncovering the relevant model's internal in a human intelligible format, without necessarily knowing its logic. Explanations can be seen as vectors of relevance scores, with the same size as the input, revealing the most important features towards a prediction.

2.3.2 Taxonomy

The categorisation of XAI methods can be done with regards to different criteria, such as specificity, type of explanation, and the approach to obtain it.

- **Intrinsic or *Post-hoc***

  XAI methods can be intrinsic by constraining the model's complexity, or *post-hoc* if the methods only affect the trained model [72]. Intrinsic models are the ones whose nature is interpretable, for example, simple decision trees or linear regressions. However, this type of method can also be obtained after applying restrictions, such as monotonicity, sparsity, or causality [11]. *Post-hoc* interpretability can only be applied in previously trained models. Nevertheless, *post-hoc* methods can be applied to intrinsic interpretable models [72].

- **Model-specific or Model-agnostic**

  Another distinction is model-specific or model-agnostic, which divides XAI methods based on their specificity to the model they are applied to. Model-specific can

only be used on particular models due to the fact that their development is based on such information.

Model-agnostic methods have no knowledge about the model as their explanations rely on pairs of inputs and outputs.

- **Saliency or Perturbation-based**

  Saliency-based methods, usually applied to visual explanations, examine gradients during the prediction to infer feature saliency. These techniques allow producing heatmaps to highlight the most important areas. Saliency-based methods are usually model-specific. On the other hand, perturbation-based methods do not know the model's internal logic, that generate explanations by perturbing the vicinity. This type of method is grounded on (1) choosing a sample to explain and how the different features are perturbed around the instance's vicinity, (2) applying the perturbed instances to the classifier and (3) the relevance or feature importance is determined by how much a perturbation has changed the classifier's prediction. Thus, a feature's perturbation that produced a severe change in the prediction implies greater relevance, but no or small change means little relevance [91]. The disadvantage is that perturbation-based methods are based on breaking the model's interactions between features during the process of perturbation, and thus, they do not consider features' dependency, which might lead to incorrect approximations in more complex models.

- **Local or Global**

  Global interpretability exists in different levels: on a holistic level, by seeking to understand how the model and all the data affects the prediction, which is very hard to obtain; on a modular level, only looking into parts of the model and their impact in the prediction [19], [72]. Local interpretability for a single prediction is based on the idea that by trying to explain a single class, it's possible to reduce the dataset, hence being more likely that the use of interpretable models can be a good approximation for obtaining interpretability. Local interpretability for a group of predictions can be achieved by using global methods on a modular level, or by applying local methods for a single prediction in all of the predictions we plan to understand [11].

2.3.3 Model-Agnostic Methods

These methods are independent of the model they are applied to considering that they retrieve *post-hoc* knowledge [92]. Figure 2.9 illustrates the working principle, through which models are treated as opaque models. Model-agnostic explanations are highly flexible as they can be applied to interpretable or black-box approaches. Their flexibility benefits, for instance, the evaluation of two different models used to predict the same

dataset [72]. Model-agnostic methods generate explanations that are not intrinsic to the classifier and thus they do not interfere with model's performance. The limitation of these models is that they are potentially less precise because they are independent of the model to be interpreted [93].

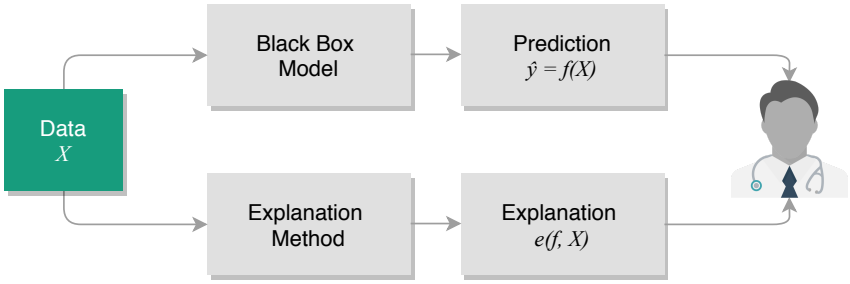


Figure 2.9: Framework for model-agnostic XAI methods. Models are treated as black-boxes, and their behaviour is explained through analysing pairs of inputs and outputs. Adapted from [11].

This dissertation focuses on several state-of-the-art model-agnostic methods, based on perturbation approaches.

2.3.3.1 Permutation Feature Importance

Permutation Feature Importance (PFI) provides global interpretability by inspecting the model score after a single feature value is randomly shuffled [94]. The increase or drop in the model score describes the relationship between the prediction and the permuted feature. Permutation Feature Importance (PFI) replaces each feature p times for other features from randomly picked instances of the dataset. Firstly, the reference score of the classifier is computed. Each feature is shuffled, generating a perturbed version of the test dataset, and the model score is recalculated using the permuted dataset. Hence, the importance of each feature is given by the average difference between the initial model score, S, and the permuted data previously repeated p times [95]. The higher the drop in the model's score, the more relevant is the feature [72].

The importance score R of each feature j is given by:

$$R_j = S - \frac{1}{p}\sum_{i=1}^{p} S_{p,j} \tag{2.9}$$

where$s_{p,j}$ corresponds to the model score when the feature, j, is permuted each p repetitions.

This method has the advantage of providing global insights about the model's behaviour, and not requiring its retrain. However, as a perturbation method, PFI assumes feature independence and does not evaluate the interaction among features.

2.3.3.2 Local Interpretable Model-Agnostic Explanations

Local Interpretable Model-agnostic Explanations (LIME), created by Ribeiro et al. [96], provide local interpretability by returning the relevance of the features for a specific instance. Portrayed in Figure 2.10, there is an example of LIME that explains mortality prediction by yielding the most relevant features. The main intuition behind LIME is that complex classifiers can be locally approximated by linear models.

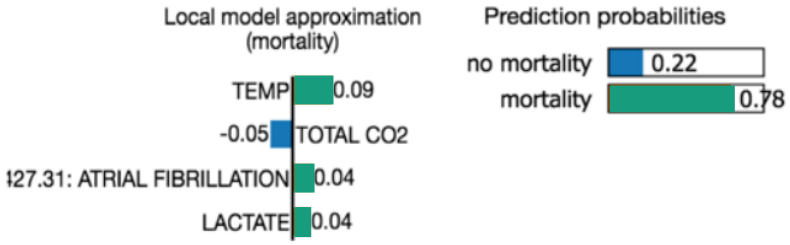


Figure 2.10: LIME example for a patient specific model interpretation. On the left, the weights of the linear regression show how the features impact the mortality prediction, where temperature, atrial fibrillation, and lactate level positively contribute to mortality prediction and total CO_2 is inversely related to mortality. On the right, LIME outputs the probability prediction from the complex model. Adapted from [97].

LIME uses a surrogate linear model to locally replace the complex model. This surrogate model is trained with the predictions given by the complex classifier, of a perturbed dataset weighted around the instance of interest. To this end, LIME approximates an interpretable model from a family of interpretable models, $g \in G$, to the complex classifier, f, while ensuring both interpretability and local fidelity by minimising the complexity of the interpretable model and the loss function, $\mathcal{L}$, respectively. Reducing the loss function reinforces the reliability of the linear local approximation to the complex classifier.

The interpretable model learns over the perturbed dataset and the respective labels. A linear model is used and weights are learned via a least-squares procedure, as showed in Figure 2.11. The linear regression weighted coefficients denote the relative importance of each feature. Higher the coefficients, the higher the impact on the prediction.

The explanation of the instance X, produced by LIME, is obtained according to Equation 2.10:

$$R = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_X) + \Omega(g) \tag{2.10}$$

$\mathcal{L}(f, g, \pi_x)$ is the loss function, which measures the unreliability of the linear approximation provided in the vicinity defined by π_X, and $\Omega(g)$ denotes the complexity of the interpretable model.

The perturbed dataset is generated in an interpretable data representation, a binary vector, which indicates the presence or absence of a given element. Given $X \in \mathbb{R}^d$ in its original representation with d dimensions, the perturbed sample is denoted as $z' \in \{0,1\}_{d'}$. Perturbations occur through the random attribution of 0 in different features.

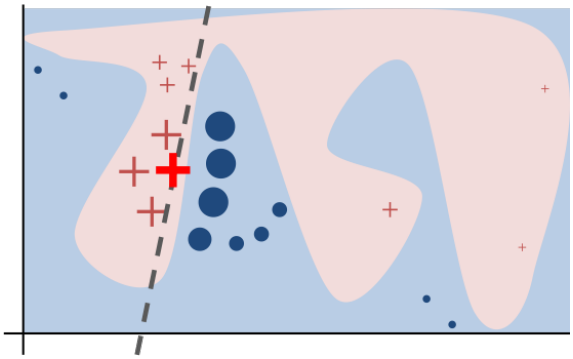


Figure 2.11: Intuition behind LIME, the complex model is illustrated by the blue and pink backgrounds, and the instance seeking to be explained is represented by the red bold cross. The additional instances are the perturbations generated, to train the linear model, the grey dash. Adapted from [92].

This dataset is transposed into the original representation, X', to draw the predictions from the complex classifier, and save them as labels. For instance, in text classification, 0 denotes the absence of a word, in image data, represents the absence of a contiguous patch of similar pixels, a super-pixel, and on tabular data, continuous features are discretised and mean centred. An illustration of several perturbations applied to image data and their respective representation in the interpretable domain is portrayed in Figure 2.12.

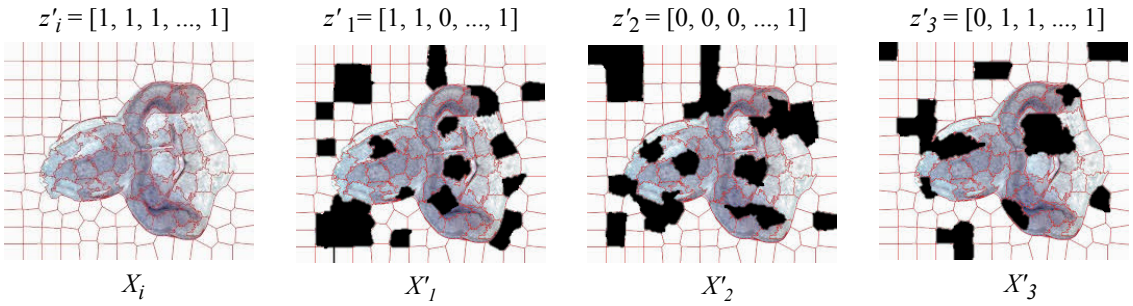


Figure 2.12: Interpretable domain examples translated into the original domain, generating several instances of the perturbed dataset, from the Drosophila discs. On the left is the instance of interest segmented according to the superpixels and on the right several random perturbations. Adapted from [98].

The perturbed instances are weighted according to Equation 2.11, through which π_X, an exponential kernel, attributes higher weights to instances similar to X:

$$\pi_i = e^{-\left(\frac{d}{\sigma}\right)^2} \tag{2.11}$$

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. Kopper and Molnar addressed the importance of defining an adequate value of σ to ensure an adequate approximation [99].

For every instance explained, it is possible to retrieve the R^2 coefficient, a measurement for how good is the linear model in approximating the complex model, i.e., how well can the linear model describe the complex predictions. This parameter is dependent on the kernel's width and the complexity of the complex models. Often more complex models, have lower R^2 values as the surrogate model fails to represent it [72].

Notwithstanding the fact that LIME provides a flexible approach to explain the prediction of every classifier, its quality is still dependent on unstudied variables, as it happens with the kernel's width and size of the perturbed dataset [72]. Further, the work of Alvarez-Melis and Jakkokola [91] on LIME reported the often existence of instability amongst explanations, where very similar instances can be provided with very different explanations.

2.3.3.3 Shapley Additive Explanations

SHapley Additive exPlanations (SHAP) is a method proposed by Slundberg et al. [63], that explains a prediction using the Shapley values, by computing the contribution of each feature.

Shapley values are derived from the literature of game theory. Parallel to a game, features in a prediction can be seen as the players and the prediction as the payout. Thus, the Shapley value of a feature or player can be described as its contribution to the final output. SHAP combines both Shapley values and LIME by including an additive feature attribution method, a linear model, into the classic perspective of Shapley values. Similarly to LIME, this method works on a simplified feature's domain, equivalent to the interpretable data representation, in which only the present features, i. e., $z'_j = 1$, play a part the classification. Based on additive feature methods, the explanation of a prediction is given by the sum of all the features' effects in the classification, approximating to the complex classifier's output [63]. The relevance score of the feature j is a linear function that can be defined by:

$$R_j = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{2.12}$$

whereϕ_i corresponds to the Shapley values associated to the ith feature on an instance.

Kernel SHAP is the model-agnostic approximation method and follows the same principle presented in Equation 2.10. It differs from LIME given that it does not calculate the parameters heuristically, and assumes the local accuracy and consistency. The local accuracy, given by Equation 2.13, assumes that the local approximation prediction matches the complex model's prediction, Consistency, presented in Equation 2.14, requires that data perturbation is reversible. The assumed properties arrive at different parameters, which consequently lead to Shapley values calculation.

$$\Omega(g) = 0 \tag{2.13}$$

$$\pi_{X_i} = \frac{(M-1)}{\binom{M}{|z|}|z|(M-|z|)} \tag{2.14}$$

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.

SHAP has a strong theoretical foundation on game theory, and by combining both LIME and Shapley values becomes more intuitive. However, Kernel SHAP is computationally expensive and implies the assumption of feature independence.

2.3.4 Explanation Quality Assesment

The assessment of XAI methods, in Figure 2.13, has three levels of explanation metrics, as proposed in the literature:: application-grounded, human-grounded, and functionally-grounded, all of them with different costs and needed resources [19].

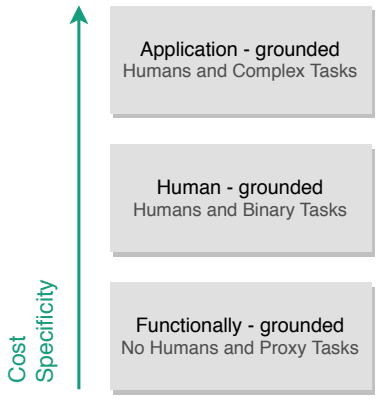


Figure 2.13: Taxonomy of XAI evaluation, from the most expensive, application-grounded, to the least expensive, functionally-grounded. Adapted from [19].

Application-grounded validation demands the performance of expert direct tasks. This validation requires the highest cost and time, as it seeks for domain experts. For instance, in the medical domain doctors or technicians are required.

Human-grounded experiments rely on simpler tasks, such as binary choices. These binary tasks can either be by forward simulation, by asking the user which is the model's output when presented with its input and explanation, or counterfactual simulation, when the user knows the input and output and is asked to change the explanation for a better fit.

Finally, functionally-grounded evaluations, are the least expensive and the first to be assessed, as they do not require humans and only depend on proxy tasks. The challenge is to define a formal definition of interpretability to be used as a proxy for explanation's quality.

In this dissertation, functionally-grounded validation methods are approached, to assess the quality of the explanations produced by the different used XAI methods. It is

suggested that an explanation must be faithful and reliable to the model it is justifying. Faithfulness measures the consistency between the predicted explanation and real prediction, and reliability assesses how representative are the explanations to the model's behaviour and with other interpretable models trained on the same data.

Chapter 3

Methods for Interpretable Time Series Classification

This chapter describes the proposed pipeline for explaining biosignals, in particular the ECG. We start initially by introducing time series notation. Next, we propose a short taxonomy for XAI methods in time series. We also present the complete proposed pipeline, including the classifiers, explanation methods and the evaluation methods. The adaptations performed to state-of-the-art model agnostic methods will also be presented. Amongst the adaptations, the use of the signal derivative is considered to provide more reasonable classification and explanation in terms of temporal dependency.

3.1 Time Series

A time series or instance is a set of samples ordered in time with a regular sampling frequency, which can be described by the following:

$$X = \{x_1, x_2, ..., x_n\} \tag{3.1}$$

where x_i, $i \in \{1, .., n\}$ is a sample.

A dataset, composed of N instances is given by:

$$D = \{X_1, X_2, ..., X_N\} \tag{3.2}$$

which is often split into two subsets, $D = \{D_{train} \cup D_{test}\}$, one to train the classifier.

Time series register the behaviour of variables over time, and thus their analysis is a meaningful tool for understanding hidden patterns in sequential data. Time series analysis is applied to several domains, to identify the correlation between past and future samples, to highlight important characteristics, and to predict future values. In the medical context, the analysis of temporal data such as the ECG, electromyogram, or

electroencephalogram, allows inferring data patterns to help clinicians in the process of diagnosis and medical decision making.

Table 3.1 summarises the used notations by presenting a simple description for each term.

Table 3.1: Time series notation used to address the methodology.

Term	Definition
Instance or Time Series, X	Set of samples ordered by time.
Sample, x	Data point which corresponds to an observation at a specific time.
Window, W	Subsequence of an instance.

3.2 Taxonomy

Despite the several works identified in Chapter 1.3 describe the taxonomy of XAI, none of the authors focused on proposing taxonomy for time series. A preliminary conceptual categorisation of time series explanations is proposed, to enable the support of our study and to stimulate research on XAI applied to time series.

The explanations of time series classifiers consist of three categories:

- **Sample-based explanation:** classifier's predictions are explained by the impact that each sample has in a particular decision.
- **Feature-based explanation:** classifier's predictions are explained by the weights of features. These features are previously calculated using feature extraction methods across different domains, such as temporal, spectral, and statistical [100].
- **Morphology-based explanation:** classifier's predictions are justified on the relevance of the instance's morphology, by extracting visually perceived attributes, such as rising slopes and falling slopes, direction, amplitude range of slope, frequency, amongst others [101].

The three proposed types are illustrated in Figure 3.1, in which special attention is given to the sample-based explanation. At the first level, sample-based explanations are issued in raw time series. Each sample has an impact on a specific prediction that can be 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 binary and multiclass classification as it simplifies any problem using a binary mindset. The explanation represents how much has a sample contributed or contravened to the final prediction.

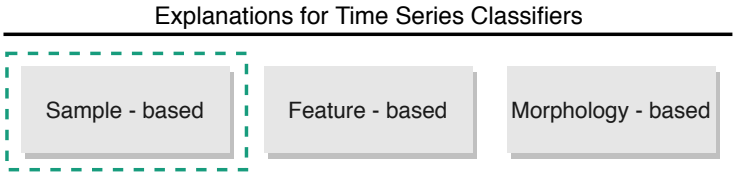


Figure 3.1: Taxonomy for time series' explanations. Amongst the three possible explanation types, the sample-based is highlighted in the image, given that it is the typology of explanation covered in this dissertation.

Model-agnostic methods for sampled-based explanations are computationally expensive and, consequently, to deal with numerous samples, a high processing time is required. Moreover, interpretability is constrained by the number of samples used in the explanation, considering that the explanation of every sample within an instance can become incomprehensible to humans. Hence it is convenient to explain a set of samples or window.

A window W is a subsequence of X and can be described by:

$$X = \{W_1, W_2, ..., W_w\} \tag{3.3}$$

through which, X can be represented as a set of windows such as $W_1 = \{x_1, ..., x_l\}$, $W_2 = \{x_{l+1}, ..., x_{2l}\}$, $W_s = \{x_{(s-1)(l+1)}, ..., x_{sl}\}$ and l is an arbitrary window size.

3.3 Model-Agnostic Methods

Several prior works addressed the problem of explaining a classifier using these methods in several applications and data types [102]–[107].

A practical example is given in Figure 3.2, where the explanation method yields the relevance scores of each instance's window.

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

$$R_i = e(f, X_i), \quad i \in \{1, ..., N\} \tag{3.4}$$

where the explanation depends on the trained classifier, f, and on the chosen instance to explain, X_i. The explanation function, e, returns the relevance scores associated to each sample, where $R_i = \{r_1, r_2, ..., r_n\}$. These relevance scores are numerical weights translating the importance of a specific sample in a decision.

3.4 Experimental Protocol

While developing a framework to explain the prediction of time series classifiers and to adapt the state-of-the-art methods, several requirements must be satisfied:

- Must explain the classifier's decision based on a set of samples with model-agnosticism;

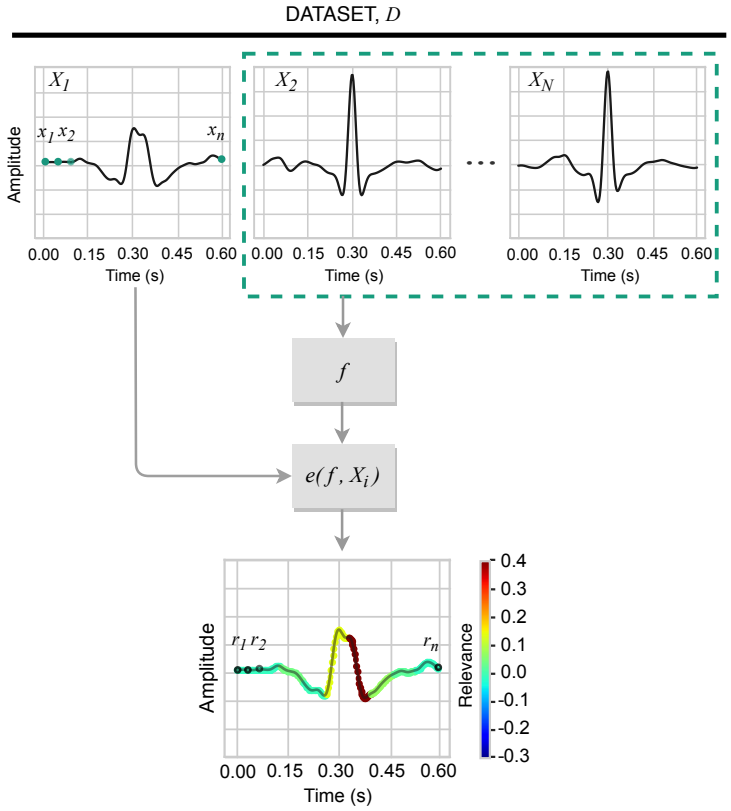


Figure 3.2: Model-agnostic methods for explaining time series. The explanation is a vector of real values that translate the relevance of each sample within the slices, with the same size as the instances from the dataset.

- Should be consistent with the prediction they are explaining;
- Should take into account sample dependency as an inherent characteristic of temporal data.

Quite often, the explanation methods produced explanations for temporal data, by calculating the relevance score of each sample in the amplitude domain. However, time series have an intrinsic one-way natural ordering, which makes dependency between features or amongst other time series an inherent characteristic. The direct application of model-agnostic and perturbation-based methods on time series overlooks temporal information as it assumes sample independence and only considers the Y-axis value. To tackle the challenge of sample independence, we argue that the introduction of the signal's derivative as a complement in the classification improves the explanation's quality. The derivative is the instantaneous rate of change, that by providing information about a sample's vicinity, considers the behaviour of a time series [108].

The impact of including the derivative in time series classification was explored by Górecki and Łuczak [108], [109], proposing a distance metric that considers the general shape of the instance, rather than just the value of a sample at a determined point in time. Additionally, the use of the derivative was employed by Keogh and Pazzani [110] and

Folgado et al. [111] as a complement to improve the alignment of Dynamic Time Warping. The mentioned works commonly reported that through the use of the derivative, higher-level features are being taken into account, 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.

An overview of the followed experimental protocol is depicted in Figure 3.3. It consists of three main stages: classify the ECG's instance using the trained model, explain the classification, and finally, evaluate the explanation's quality.

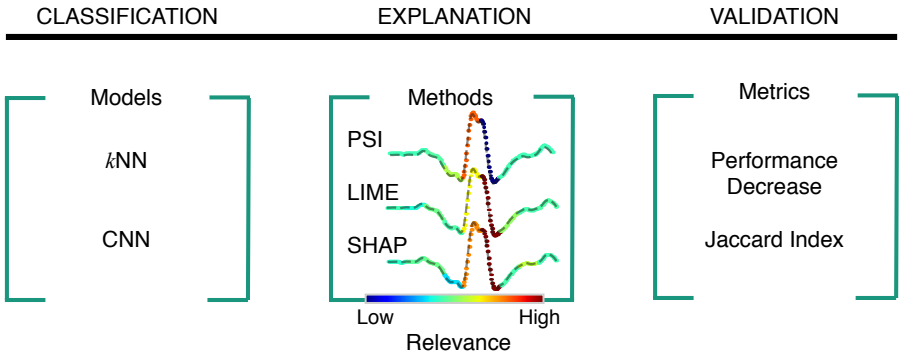


Figure 3.3: Schematic representation of the followed protocol. Three stages were considered the classification, explanation and validation.

During the experiments, two different classifiers were considered. On one hand, the interpretable k-NN and on the other, the black-box classifier, a CNN.

These models were trained with two types of data combinations:

- $Amplitude$ **:** consisting of the raw signal.
- $Amplitude + Derivative$ **:** encompassing the signal combined with its derivative.

The explanations are provided by calculating the relevance scores using the methods described in Section 3.3, namely PFI, LIME, and SHAP, on a public dataset of ECG. The evaluation of the method on whether the sample's relevance is correctly calculated follows two validation strategies: comparison with Shapelets using the Jaccard index and performance decrease.

3.5 Classification

For heartbeat classification, ECG data from MIT-BIH Arrhythmia Database requires pre processing, which includes data filtration, normalisation, and segmentation.

The ECG signal is susceptible to artefacts due to baseline wander, powerline interference, muscular interference, and electrode motion. Baseline wander is a low-frequency artefact, in the range of 0.5 Hz, related to movement and patient respiration. Powerline interference is a common noise source that has a characteristic frequency of 50 or 60 Hz.

On the contrary, muscle interference presents a large spectral content of frequency from 5 to 50 Hz, which often overlaps with ECG waves and hence being more complex to remove. The frequency associated with the cardiac cycle is around 1 to 5 Hz, thus the application of a pass band filter with a cut off value of 4 to 22 Hz, removes embedded artefacts and allows noise reduction [112]. Data acquired from different patients is associated with variable amplitude ranges of electric potential. In this sense, minimum to maximum normalisation was applied to ensure that data amplitude was within the interval $[-1, 1]$ in each heartbeat, independent of the subject and acquisition. The application of this normalisation to the recordings of each subject allows limiting the amplitude of heartbeats while maintaining the relative differences between normal and ectopic amplitudes. Data segmentation is based on the existing annotations of the dataset, which comprise the occurrence of each R peak. Each heartbeat is centred in the aforementioned peak and composed of a total of 600 milliseconds, 300 milliseconds before the annotation, and 300 milliseconds after. Combined data, $Amplitude+Derivative$, has twice the length because it includes its derivative. To compare explanations from different classifiers, a k-NN and a CNN classifier are built on the two types of data. These models were chosen due to their presence in the literature related to time series classification, whereas the k-NN stands as a simple model that makes predictions based on the comparison from the training data, and the CNN is a more complex model that enables 1-D convolution on the instances to extract information.

The k-NN model uses $k = 5$ nearest neighbours and the euclidean distance metric. The CNN architecture is slightly different for binary and multiclass since it was optimized for each case individually. The CNN binary classifier, detailed in Table 3.2, was based on Kachuee et al. [113].

Table 3.2: Description of the binary CNN architecture. Layers 1 to 6 use ReLU as activation function while layer 7 uses Softmax. F=Number of Filters, K=Kernel Size, P=Pool Size, S=Stride, U=Units and N=Number of Outputs.

Block	Layers
1	Convolution1D-(F32,K5,ReLU), MaxPool1D-(P5,S2)
2	Convolution1D-(F64,K5,ReLU), MaxPool1D-(P5,S2)
3	Convolution1D-(F64,K5,ReLU), MaxPool1D-(P5,S2)
4	Convolution1D-(F128,K5,ReLU), MaxPool1D-(P5,S2)
5	Convolution1D-(F128,K5,ReLU), MaxPool1D-(P5,S2)
6	Flatten, Dense-(U16,ReLU)
7	Dense-(N2,Softmax)

The CNN binary architecture is composed of a five block 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 with 128. These blocks are followed by two fully-connected layers with 16 neurons and a softmax layer to predict output class

probabilities. The CNN multiclass architecture, presented in Table 3.3, includes four convolutional blocks with different parameters.

Table 3.3: Description of the multiclass CNN architecture. Layers 1 to 6 use ReLU as activation function while layer 7 uses Softmax. F=Number of Filters, K=Kernel Size, P=Pool Size, S=Stride, U=Units and N=Number of Outputs.

Block	Layers
1	Convolution1D-(F156,K27,S1,ReLU), MaxPool1D-(P2,S2)
2	Convolution1D-(F64,K14,ReLU), MaxPool1D-(P2,S2)
3	Convolution1D-(F56,K3,ReLU), MaxPool1D-(P2,S2)
4	Convolution1D-(F32,K1,ReLU), MaxPool1D-(P2,S2)
5	Flatten, Dense-(U126,RelU)
6	Dense-(32,ReLU)
7	Dense-(N4,Sofmax)

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.

3.6 Explanation

In the context of sample-based explanations, the explored model-agnostic methods are based on perturbations, in which both the window's size and the number of instances in the perturbed dataset influence the explanation. These parameters should seek to maximise the meaning of the explanation and the quality of the approximation to the complex classifier.

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. It would be ideal to use a small window's length and a large number of perturbed instances. However, the computational time for such an extent dataset becomes untractable.

On the other hand, windows that are too large end up not providing useful information about the segments, only allowing a generalization of their impact. Based on exploratory analysis, we empirically defined a limited number of 1000 instances in the perturbed dataset, the use of nine windows maximised the mean scores for each slice. Therefore, the number of perturbed instances for each method was set to 1000, and the relevance scores were calculated from fixed-length windows of size of 24 samples, which corresponds to 0.07 seconds, using a total of nine windows to explain the classification.

The adaptions applied to the several XAI methods for explaining temporal data are clearly described in the following sections.

3.6.1 Permutation Sample Importance

The intuition of PFI is translated into a local XAI method. For convenience, we denote the method Permutation Sample Importance (PSI), emphasizing that is applied to the raw time series and not directly on the features extracted from time series.

To provide explanations for a single prediction at the time, the permuted dataset, $\tilde{D}$, is composed of several copies of the time series to be explained, with the samples permuted. The permutations are applied to each sample within a window randomly chosen from the time series. The permutation replaces the samples' value with other belonging to the opposite class or classes. This step forces the new instance to have permutations from a different class than the one intended to be explained. The intuition behind this replacement is to generate a perturbed instance, different enough to modify the classifier's prediction.

The relevance of a sample from the instance i, r, is defined as the mean of the differences between the *a posteriori* probabilities of the baseline and permuted dataset, as defined according to Equation 3.5:

$$r = \frac{1}{p} \sum_{j=1}^{p} P(\hat{y}_i|D) - P(\hat{y}_i|\tilde{D}) \tag{3.5}$$

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

For the PSI three permutations for each slice (p) are performed. The explanations for each instance in the test set are calculated, generating a matrix of the same size as D_{test}, used in the next stage of the protocol.

3.6.2 Local Interpretable Model-agnostic Explanations

LIME was initially proposed to explain image, tabular and text data, thus the original idea is adapted into explaining time series. Instead of perturbing features, the perturbation applies to the time series' windows. The perturbed window $\tilde{W}$ are given as follows:

$$\tilde{W}_w = \begin{cases} W_w, & \text{if } z'_w = 1 \\ p(W_w), & \text{if } z'_w = 0 \end{cases} \tag{3.6}$$

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

The deletion of samples in time series, when $z_w = 0$, can not be translated into removing or replacing with missing values, as most machine learning algorithms do not support data with missing values. To reproduce the deletion, there are numerous possible perturbation functions, and each results in different explanations. Three perturbation functions, defined in Equations 3.7 to 3.9, were considered. Examples of those perturbations are illustrated in Figure 3.4.

The **zero** perturbation is defined as:

$$p(W_w) = 0 \tag{3.7}$$

The **random** perturbation is defined as:

$$p(W_w) = W_w + \theta\mathcal{N}(0,1) \tag{3.8}$$

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

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

$$p(W_w) = \frac{1}{Xl}\sum_{i=1}^{X}\sum_{j=1}^{l} X_i[(j-1)(l+1), jl] \tag{3.9}$$

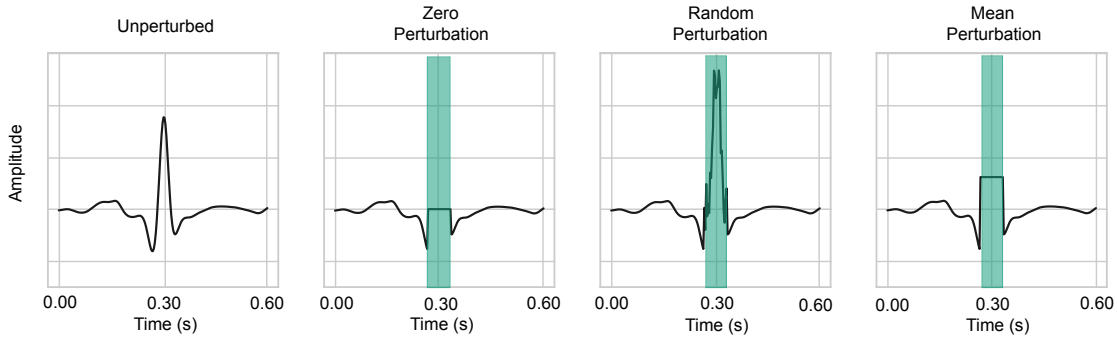


Figure 3.4: Example of the three possible perturbations applied to the R peak of the heartbeat representation. On the left, the unperturbed instance for comparison, followed by the zero perturbation, and random perturbation, and finally, on the right, the mean perturbation.

The faithfulness of LIME in describing the complex classifier determines if the linear model is faithful to the classifier, i.e. if it correctly approximates 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 correct local prediction with a high R^2 implies that the linear model is correctly approximating the complex model for a given instance. To determine the best fitting perturbation, faithfulness is analysed in Section 4.2.2.

3.6.3 Shapley Additive Explanation

As addressed in Section 2.3.3.3, Kernel SHAP determines the relevance of each sample within the instance of interest. To perturb an instance, Kernel SHAP applies a mask function that contains vectors in the interpretable domain representation. 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.

Kernel SHAP is modified to obtain the contribution of each window in the prediction, instead of each sample. With this purpose, the mask function was adapted, to force the absence, or presence, of all the values within a window.

3.7 Validation

To validate the explanations, the most relevant window is replaced to evaluate the classifier's response and to compare them against the discriminative subsequences of each class. Since the methods used 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, we considered the most relevant window of each instance, i.e., the one with the maximum relevance score within that explanation.

3.7.1 Jaccard Index

The 2-D Jaccard index is often used in computer vision to evaluate image segmentation and object detection algorithms [114], [115]. It is proposed that the Jaccard index in 1-D is used to compare the most relevant windows calculated using the adapted methods described, with Shapelets extracted from the time series.

Given that Shapelets are subsequences that maximally describe a given class, the similarity between the Shapelets against the most important window can be a means of determining the explanation's quality.

The similarity is measured as defined in Equation 3.10, outlined in Figure 3.5.

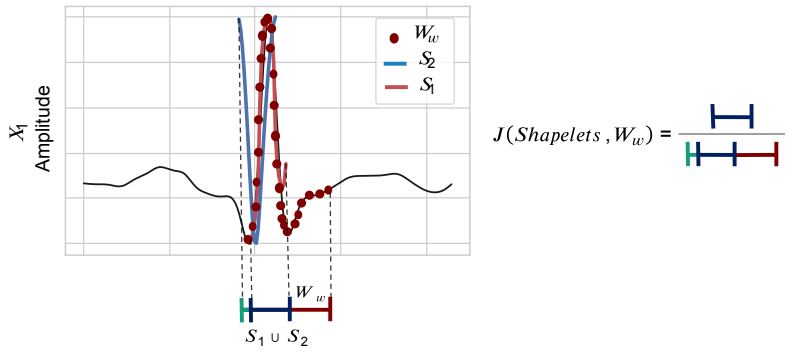


Figure 3.5: Example of 1-D Jaccard's index calculation trough the comparison of sets extracted from the Shapelets classifier and the most relevant subsequences determined by the explanation method, where $Shapelets = S_1 \cup S_2$.

$$J(Shapelets, W_w) = \frac{\#Shapelets \cap W_w}{\#Shapelets \cup W_w} \tag{3.10}$$

$J(Shapelets, W_w)$ 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.

To determine the Shapelets and their predicted locations for minimum distance the learning Shapletes algorithm from tslearn Phyton package [116] were implemented. For every explanation matrices from the different methods, the Jaccard index is determined comparing the location of the most relevant window from the explanation, against one discriminative subsequence from the learning Shapelets. In order to be comparable, the Shapelets classifier was trained with one shapelet with the same size of the segments used to produce explanations.

3.7.2 Performance Decrease

The quality of explanations can be assessed by analysing models' performance. In particular, the concept of performance decrease provides information about a given explanation as it replaces the most relevant windows and recalculates the classifier's performance. The performance decrease is the difference between the performance of the classifier with the unaltered data and after the most relevant window on the dataset is replaced. Slight changes or increases in the performance decrease indicate that the explanations are not representative of the model.

Different replacement methods are considered, such as zero, inverse, and swap, according to the work of [67]. The methods are applied to the windows W_w with relevance r_w equal δ, the maximum relevance score within R:

$$W'_w = \begin{cases} W_w, & \text{if } r_w < \delta \\ v(W_w), & \text{if } r_w \geq \delta \end{cases} \tag{3.11}$$

where v is the replacement function.

The replacement methods are defined by Equations 3.12 to 3.14.

The **zero** substitution is defined as:

$$v(W_w) = 0 \tag{3.12}$$

The **inverse** substitution is defined as:

$$v(W_w) = \max(X_i) - W_w \tag{3.13}$$

The **swap** substitution is defined as:

$$v(W_w) = \{x_{m+k}, x_{m+k-1} ..., x_m\} \tag{3.14}$$

The zero substitution is similar to the perturbation outlined in Equation 3.7. The reverse and swap substitutions rely on symmetry about the amplitude and the time axis, respectively.

The different substitutions considered, carry distinct interpretations. The swap substitution is particularly relevant in the context of time series due to the fact that it performs the replacement with the same subsequence but in an opposite temporal ordering. If swapping the samples in their opposite temporal ordering do not impact negatively the model's performance, it might convey that the sample's temporal dependency is not being taken into account.

Chapter 4

Results

This chapter outlines the results obtained by applying the methods discussed in the previous chapter to ECG data. Firstly, the dataset to validate the proposed methods is detailed. Thereafter, the results which relate to the predictive performance of the algorithms, and the validation of the explanations calculated using the proposed methods are presented. The results are reported divided into binary and multiclass classification.

4.1 Dataset Description

The MIT-BIH Arrhythmia Database [117] 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. According to the Association for the Advancement of Medical Information (AAMI) practices [118] the classified beats are organised in normal (N) and ectopic classes (E), the latter further divided into ventricular ectopic beat (V), supra-ventricular ectopic beat (S), and unknown (Q).

To compose the train and test datasets, in the clinical context, data must be divided by patient rather than by individual heartbeat, since ECG recordings within the same subject are highly correlated between them. The dataset is split into train and test according to De Chazal et al. [119], which guarantees that the patients in the test set were not used during training. The specific composition of each dataset is detailed in Table 4.1.

Splitting the data allows to accurately evaluate the model and prevents overfitting,

caused when the model is adjusted to the training data but is not able to make generalisations on new data.

Table 4.1: Composition of train and test dataset, according to the subjects present in MIT BIH-Arrythmia Database.

Dataset	**Subjects**
Train	101, 106, 108, 109, 112, 114, 115, 116, 118, 119,122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, 230
Test	100, 103, 105, 111, 113, 117, 121, 123, 200, 202,210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, 234

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 [120].

This dataset is highly unbalanced as it can be observed in Table 4.2, which makes the classification a challenging task.

Table 4.2: Class distribution of MIT-BIH arrhythmia database heartbeat types into the AAMI heartbeat classes, in the train and test set. The considered classes for this work followed the AAMI standards: N, S, V and F.

AAMI Heartbeat Classes	**MIT BIH-Arrythmia Database**	**Train**	**Test**
Non-Ectopic beats (N)	Normal beat Left bundle branch block beat Right bundle branch block beat Nodal escape beat Atrial escape beat	45805	44019
Supraventricular ectopic beats (S)	Aberrated atrial premature beat Premature or ectopic supraventricular beat Atrial premature contraction Nodal escape beat	975	2049
Ventricular ectopic beats (V)	Ventricular flutter wave Ventricular escape beat Premature ventrcular contraction	3788	3220
Fusion beats (F)	Fusion of ventricular and normal beat	414	388

To adjust the class distribution, further random undersampling was applied in the training data. Random undersampling is a technique to rebalance the distribution of classes of unbalanced datasets, through the random removal of instances. In the multiclass case, instances from the two most prevalent classes, N and V of the training dataset, were removed, totalling a distribution of 975 N heartbeats and 975 S heartbeats. In the binary dataset, the instances of the training set from the N class were removed, remaining 5177 N heartbeats.

4.2 Binary Classification

First, the initial task focused on binary classification, to distinguish between normal and abnormal beats, also referred to as ectopic beats. The results for the binary classification are presented in the following sections.

4.2.1 Classifiers

Given that the dataset is highly unbalanced, the classifiers' performance in both the binary and multiclass cases is not reported using accuracy but instead assessed by measuring the recall, precision and F_1 score.

Table 4.3 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 4.3: Model's performance in binary classification. The two classes are 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 4.3 shows evidence that the overall performance of the CNN 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 a decrease in the F_1, recall, and precision scores. Although the classifiers that only use the amplitude obtain better performances, the derivative component might lead towards more reasonable models and explanations in terms of temporal dependency between samples. Therefore in the following sections, the quality of the explanations with both approaches is assessed.

4.2.2 Faithfulness

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 and the fitting of the linear model to the complex model is represented through the R^2. Table 4.4 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, which are consistently higher for the mean and zero in comparison with Random.

Table 4.4: Faithfulness of LIME measured by means of F_1, recall, precision scores and the mean R^2(standard deviation), according to the different possible substitutions, zero, random and mean, to produce the explanations.

	***k*-NN** $_{Amp}$			***k*-NN** $_{Amp+Dev}$		
	Mean	Zero	Random	Mean	Zero	Random
F_1	91.9	**92.4**	89.1	95.2	93.1	**95.7**
Recall	73.6	76.8	**85.0**	**96.0**	91.1	94.3
Precision	**97.6**	95.5	77.2	91.0	89.7	**93.7**
R^2	0.61 (0.27)	**0.67 (0.20)**	0.27 (0.12)	**0.55 (0.22)**	0.47 (0.23)	0.20 (0.19)
	CNN $_{Amp}$			**CNN** $_{Amp+Dev}$		
	Mean	Zero	Random	Mean	Zero	Random
F_1	**98.9**	97.1	98.5	**97.8**	95.3	87.8
Recall	97.6	**98.5**	97.7	95.3	96.2	**99.9**
Precision	**96.1**	85.6	93.3	**97.9**	94.4	71.7
R^2	0.59 (0.20)	**0.66 (0.16)**	0.57 (0.20)	**0.67 (0.17)**	0.56 (0.14)	0.40 (0.10)

Both the mean and zero substitution methods present similar faithfulness. The mean substitution was chosen to be used in evaluating the quality of the explanations provided by LIME for the binary classification. The results also indicate that the local approximation provided by LIME is adequately approximating the complex model.

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.

4.2.3 Jaccard Index

Jaccard's index is used as a metric to evaluate the explanation. As discussed in Section 3.7.1 we used the Jaccard index to measure 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 4.5 summarises the results for the three explanation methods considered for the considered classifiers. For baseline comparison, an additional explanation method, denoted as Random, was included to assign random relevance scores.

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

Table 4.5: 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, which presents the average Jaccard index (standard deviation).

	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)			

In general, 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. [67], which reported identical low overlap between the Shapelets and the XAI method when using less than two Shapelets. In the absence of a ground-truth for the explanations, the Shapelets were considered as an approximation, considering that they theoretically represent the most discriminative region of each segment. In this case, the Jaccard's index is being deployed by assuming the Shapelets-based classifier as a ground-truth. However, that is not entirely real 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. However, its particular use in time series explanations when comparing it to the Shapelets still raises the question of which method could be used for comparison as the ground-truth.

4.2.4 Performance Decrease

The previous section presented an analysis regarding the agreement between the explanation methods and the Shapelets, as a baseline method to retrieve the most relevant subsequences towards the classification. In this section, it is presented 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 4.6 shows the decrease in the F_1 score when the perturbations presented in Section 3.7.2 are applied, for the binary task.

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. Due to the fact that 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.

Table 4.6: F_1 score's decrease of the binary classification. 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	3.1	-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

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 more sensitive to its temporal ordering.

Contrarily, SHAP presents values close or below the performance decrease of random relevance scores. The low performance of SHAP was related to attributing high relevance scores to the last window as shown in Figure 4.1, in which are presented various explanations for correct classifications of normal beats.

Often the last window has greater relevance, which produces explanations that are not consistent with the classifier, meaning that the perturbations of that slice do not impact the model's performance when compared to the random weights. Therefore, both the Jaccard index and the performance decrease present results below the envisioned. In the case of 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. The results for LIME and PSI, using the k-NN, show that adding the derivative brings temporal information into the explanation. On the contrary, 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.

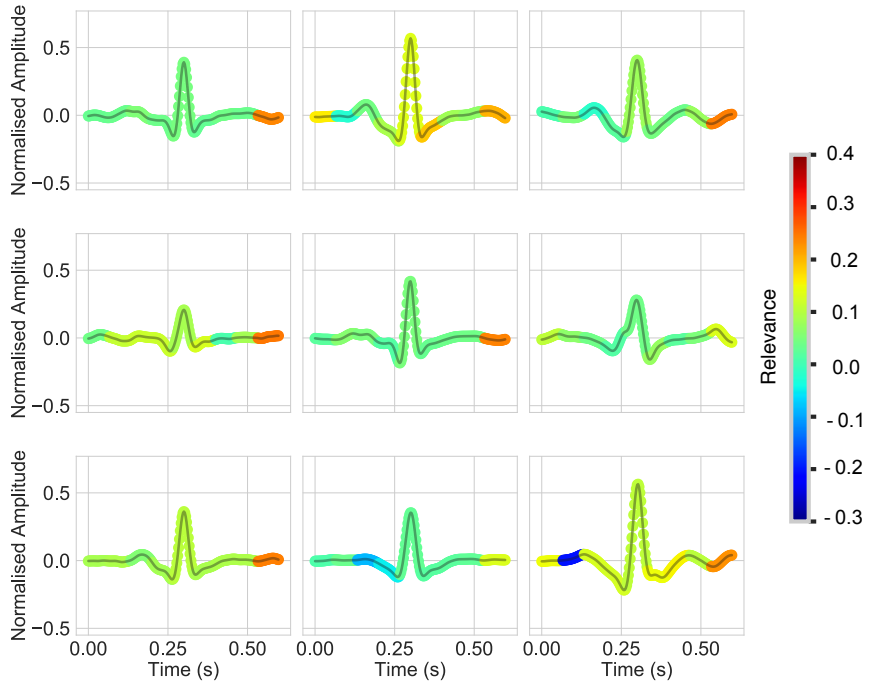


Figure 4.1: Representation of SHAP explanations for several predictions of the binary $\text{CNN}_{Amp+Dev}$. The most relevant subsequence usually occurs for the last window.

4.3 Multiclass Classification

The multiclass case results are detailed in the following sections, which comprise the performances' of the multiclass classifiers and explanations' results.

To reduce computational time, it was assumed that the values of faithfulness were within the same range and the mean substitution was also chosen to produce the LIME explanations.

4.3.1 Classifiers

Table 4.7 summarises the classifiers' performance. 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 $\text{CNN}_{Amp+Dev}$.

Table 4.7: 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**	**Av**	**N**	**S**	**V**	**F**	**Av**	**N**	**S**	**V**	**F**	**Av**
***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

4.3.2 Jaccard Index

The results for the multiclass classification are presented in Table 4.8. 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%.

Table 4.8: 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, which presents the average Jaccard index (standard deviation).

	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)			

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.

In general, the Jaccard indexes are higher in the multiclass case in comparison to the binary case. The increase in the similarity between the location of the Shapelets and the most relevant window can be explained by less variability and more specificity amongst the time series of the same class when considering the three ectopic classes separately.

In the binary case, the abnormal class presents characteristics that are specific to several anomalies. In particular, the S class, supraventricular ectopic beats, are characterised by extrasystole and premature beats that occur prior to the QRS complex, while the V class, ventricular ectopic beats, present a wider QRS complex that occurs earlier than expected. Thus it is more complex for a local explanation to be compatible with the behaviour of the Shapelet-based model, than in the multiclass case.

4.3.3 Performance Decrease

Table 4.9 presents the performance analysis for the multiclass case. 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. 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.

Table 4.9: 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

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. For the multiclass cases, the agnostic models convert the task into a binary situation and return the relevance of each window into reaching the prediction of the complex classifier. Each explanation generated is binary, in the sense that positive scores represent the positive contribution in such decision and negative scores represent the contribution towards the opposite classes. Therefore, the explanation is more specific to the classifier and has greater quality.

Figure 4.2 represents various explanations for correct classifications of a normal heartbeat from the multiclass k-NN$_{Amp+Dev}$ with SHAP.

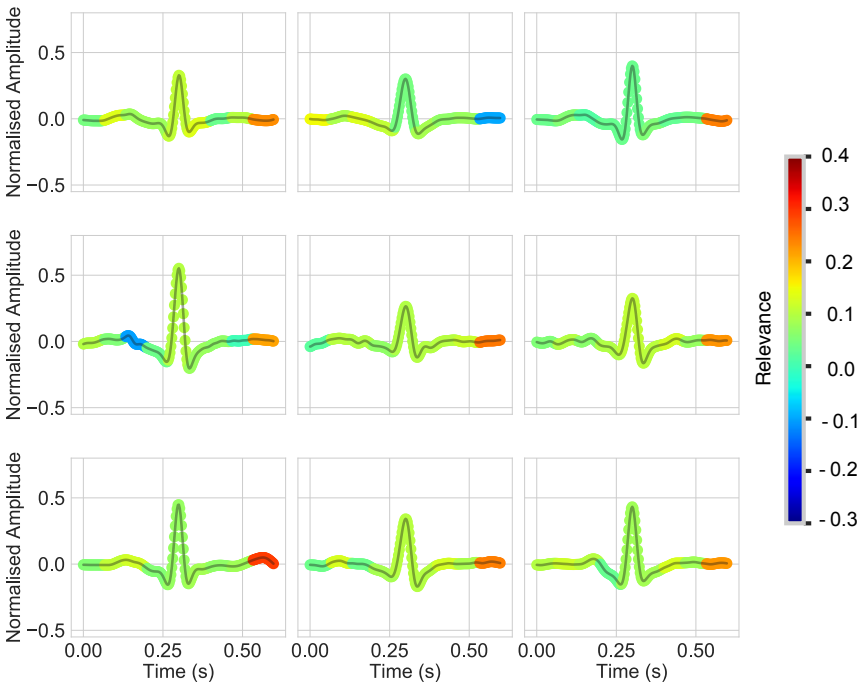


Figure 4.2: Representation of SHAP explanations for several predictions of the multiclass KNN$_{Amp+Dev}$. The most relevant subsequences usually correspond to the last window.

The explanations provided are similar to the binary classifier, demonstrating high

relevances for the final ECG segments. The explanations for SHAP in the multiclass case, show poor results when included the derivative.

In the literature, SHAP is reported as a reliable agnostic method for explaining predictions [64], [103]. 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.

4.4 Use Case on Explaining Misclassifications

The results presented in the last sections 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.

The debugging of a machine learning model is often a complex task, that can be made easier by further understanding why a model is misclassifying instances. A preliminary approach to validate the XAI methods relies on understanding whether the explanations allow gaining insights into the malfunction of the model and if the provided information is effective to increase the quality of the model. Figure 4.3, illustrates several explanations for correct S class predictions for the multiclass CNN_{Amp}.

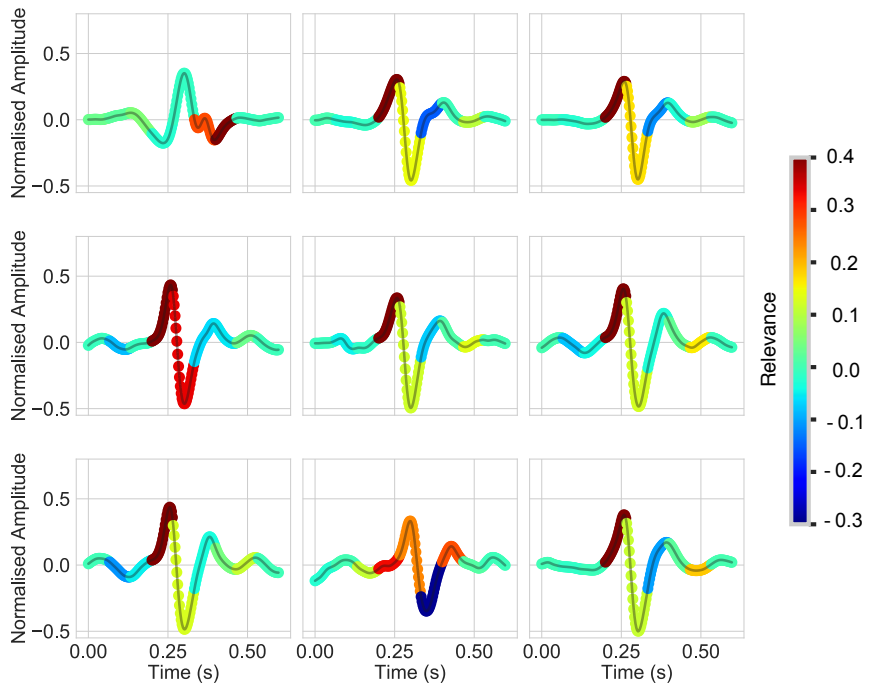


Figure 4.3: Representation of the CNN_{Amp} behaviour. LIME explanations for correct classifications of the S class.

Figure 4.3 demonstrates that the QR segment presents the greatest relevance into the

decision of supraventricular ectopic beats for the model in question. In contrast, Figure 4.4 presents some of the false negatives for the S class that were incorrectly classified as normal.

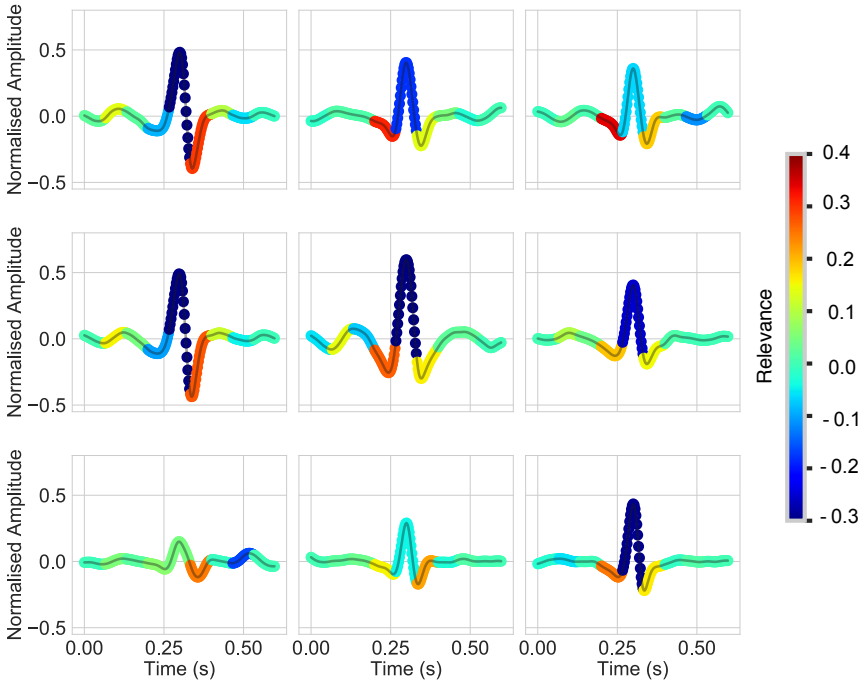


Figure 4.4: Representation of the CNN_{Amp} misbehaviour. LIME explanations for false negatives of class S.

The heartbeats in question were classified as normal mostly due to the Q and S peaks, yet the R peak contradicts the model's classification. 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.

The further realignment of the misclassified heartbeats has shown to impact the model's performance. Shifting the signal by 0.04 seconds, allowed to correct the previously incorrect classifications of class S beats by 40%. Through the explanations for the misclassifications, it was possible to understand the model's weaknesses and correct them.

Chapter 5

Conclusion and Future work

This chapter reports the main contributions and conclusions of this dissertation, followed by several suggestions for future research.

5.1 Conclusion

The lack of interpretability of top-line performance models has hampered the acceptance of AI in the medical field. Additionally, there is no consensus protocol to assess the explanation's quality, without human intervention. Furthermore, model-agnostic explanations for time series classification are still preliminary and are a challenging task, as they often assume feature independence, which resembles in the time series use case, temporal independence between samples.

In this dissertation, a preliminary taxonomy for time series explanations is presented and the sample-based explanations are addressed by the adaptation of several model-agnostic XAI methods into explaining the relevance of each window for the classification of a given instance. Hence, the main contributions consist of an extensive evaluation of several explanation methods applied to the time series in the MIT BIH Dataset and the proposal to use of the derivative as a method to design models that perform well and capture temporal explanations in data. The derivative introduces the notion of temporal dependency to the explanation as it is, by definition, the instantaneous rate of change of a signal. Lastly, the proposed adaptation is validated against the other model-agnostic methods, using a public ECG dataset.

To fairly assess the proposed framework, the dataset was approached in two different means, according to binary, and multiclass classification. The faithfulness of each substitution is a relevant topic when dealing with LIME, a model-agnostic perturbation-based method to explain time series. Therefore, in the binary case, the reliability of each local

prediction by LIME was evaluated, denoting that the mean and zero substitution present similar faithfulness.

For validating the explanations produced by the different methods, it was analysed 1) the Jaccard Index, which took into account the Shapelets model as a reference, and 2) the performance decrease. The Jaccard index demonstrated a higher similarity between the Shapelets model and the explanation for the multiclass situation when in comparison to the binary classification, caused by a more specific separation of classes. Although the average Jaccard index still requires further research about what information to be used as ground-truth, it demonstrates a promising and feasible application. Regarding the analysis of the models' performance it can be concluded that for models with simpler internal logic, the use of the derivative is beneficial as it adds information about the samples' dependency, for explanations calculated using LIME and PSI.

After the results, the importance of explanation methods was demonstrated through a concrete use case. In particular, the explanation methods made it possible to assess the reason for the misclassification and were useful in understanding the model's behaviour and improving it.

Overall, this dissertation aims to provide a basis for reliable explanations and inspire future research on reusing the proposed methodology for real application scenarios.

5.2 Future Work

The developed work shows promising results, unveiling the advantages of introducing the signal derivative to include temporal information into the sample-based explanations, particularly in LIME and PSI.

The XAI methods approached are computationally expensive as they rely on the further classification of the perturbed synthetic dataset to obtain explanations. The high processing time suggests the possible future optimisation of these methods. In this dissertation, the first level of explanation, i.e., sample-based explanation, proposed in the initial time series taxonomy was explored. The remaining levels, in particular the feature-based and morphology-based explanations, have been left for future work.

Throughout this dissertation, a fixed-length window was used to produce explanations. However, another pertinent research would consist of using automatic segmentation, to define variable-length windows that properly explain segments based on the signal morphology, instead of the fixed arbitrary length. The variable segmentation along the signal would rely on the notable events of the ECG and could increase the flexibility of the explanation and provide more meaningful information.

Moreover, although quantitative methods for validating explanations are explored, when considering the clinical case, qualitative methods that include validation with humans are needed. Hence, the validation protocol could include human interaction to assess the explanations, and to be used as ground-truth, instead of the Shapelets-based

model. In an initial phase, validating with non-experts, by questioning users with simple tasks and its explanations, and later on, using expert domains, such as doctors.

Finally, it would be relevant to combine all types of explanations for time series, in an interface, that would resemble human explanation, a complex set of various information, based on the feature, the morphology and the sample.

Bibliography

[1] World Health Organization, *Cardiovascular diseases (CVDs)*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (visited on 08/13/2020).

[2] M. Sharma, R. S. Tan, and U. R. Acharya, "Automated heartbeat classification and detection of arrhythmia using optimal orthogonal wavelet filters," *Informatics in Medicine Unlocked*, vol. 16, no. May, p. 100 221, 2019, ISSN: 23529148. DOI: 10.1016/j.imu.2019.100221. [Online]. Available: https://doi.org/10.1016/j.imu.2019.100221.

[3] S. Schneiter, L. D. Trachsel, T. Perrin, S. Albrecht, T. Pirrello, P. Eser, B. Gojanovic, A. Menafoglio, and M. Wilhelm, "Inter-observer agreement in athletes ECG interpretation using the recent international recommendations for ECG interpretation in athletes among observers with different levels of expertise," *PLoS ONE*, vol. 13, no. 11, 2018, ISSN: 19326203. DOI: 10.1371/journal.pone.0206072. [Online]. Available: /pmc/articles/PMC6248914/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6248914/.

[4] D. Massel, "Observer variability in ECG interpretation for thrombolysis eligibility: Experience and context matter," *Journal of Thrombosis and Thrombolysis*, vol. 15, no. 3, pp. 131–140, 2003, ISSN: 09295305. DOI: 10.1023/B:THRO.0000011368.55165.97. [Online]. Available: https://link.springer.com/article/10.1023/B:THRO.0000011368.55165.97.

[5] Rounak, *AI and Machine Learning Impact The Healthcare Industry | Opencodez*. [Online]. Available: https://www.opencodez.com/how-to-guide/how-can-artificial-intelligence-and-machine-learning-impact-the-healthcare-industry.htm (visited on 01/12/2020).

[6] J. Ordish, H. Murfet, and A. Hall, *Algorithms as medical devices Acknowledgements*. PHG Foundation, 2019, ISBN: 9781907198335. [Online]. Available: www.phgfoundation.org.

[7] W. Fan, J. Liu, S. Zhu, and 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*, pp. 1–26, 2018,

ISSN: 15729338. DOI: 10.1007/s10479-018-2818-y. [Online]. Available: https://doi.org/10.1007/s10479-018-2818-y.

[8] T. Sullivan, *Half of hospitals to adopt artificial intelligence within 5 years | Healthcare IT News*, 2017. [Online]. Available: https://www.healthcareitnews.com/news/half-hospitals-adopt-artificial-intelligence-within-5-years (visited on 06/16/2020).

[9] T. Q. Sun and R. Medaglia, "Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare," *Government Information Quarterly*, vol. 36, no. 2, pp. 368–383, 2019, ISSN: 0740624X. DOI: 10.1016/j.giq.2018.09.008.

[10] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 35–43, 2018, ISSN: 15577317. DOI: 10.1145/3233231. arXiv: 1606.03490.

[11] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics (Switzerland)*, vol. 8, no. 8, pp. 1–34, 2019, ISSN: 20799292. DOI: 10.3390/electronics8080832.

[12] P. Gandhi, *Explainable Artificial Intelligence*. [Online]. Available: https://www.kdnuggets.com/2019/01/explainable-ai.html (visited on 01/09/2020).

[13] F. Cabitza, D. Ciucci, and R. Rasoini, "A giant with feet of clay: On the validity of the data that feed machine learning in medicine," 2017.

[14] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, 2019, ISSN: 1942-4787. DOI: 10.1002/widm.1312. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1312.

[15] A. Burt, *Is there a 'right to explanation' for machine learning in the GDPR?* [Online]. Available: https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/ (visited on 01/13/2020).

[16] R. Hamon, H. Junklewitz, and I. Sanchez, "Robustness and explainability of artificial intelligence - From technical to policy solutions," Tech. Rep., 2020, p. 40. DOI: 10.2760/57493. [Online]. Available: https://dx.doi.org/10.2760/57493.

[17] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI," 2019. arXiv: 1907.07374. [Online]. Available: http://arxiv.org/abs/1907.07374.

[18] S. Dreiseitl and M. Binder, "Do physicians value decision support? a look at the effect of decision support systems on physician opinion," *Artificial intelligence in medicine*, vol. 33, pp. 25–30, 2005. DOI: 10.1016/j.artmed.2004.07.007.

[19] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," 2017. arXiv: 1702.08608. [Online]. Available: http://arxiv.org/abs/1702.08608.

[20] Johner Institute, *Artificial Intelligence AI for Medical Devices - A Regulatory Perspective*. [Online]. Available: https://www.johner-institute.com/consulting/technical-documentation/ai-medical-devices/ (visited on 01/22/2020).

[21] E. Bobek and B. Tversky, "Creating visual explanations improves learning Cognitive Research: Principles and Implications," *Johnstone*, vol. 1, p. 27, 1994. DOI: 10.1186/s41235-016-0031-6.

[22] S. Liu, X. Wang, M. Liu, and J. Zhu, "Towards better analysis of machine learning models: A visual analytics perspective," *Visual Informatics*, vol. 1, no. 1, pp. 48–56, 2017, ISSN: 2468502X. DOI: 10.1016/j.visinf.2017.01.006. arXiv: 1702.01226.

[23] O. Biran and C. Cotton, "Explanation and Justification in Machine Learning: A Survey," *IJCAI Workshop on Explainable AI (XAI)*, no. August, pp. 8–14, 2017.

[24] B. A. M. Turing, "Computing machinery and intelligence," *Mind, New Series*, vol. 59, no. 236, 1950.

[25] D. Crevier, *AI: The Tumultuous History of the Search for Artificial Intelligence*, April. BasicBooks, 1993, ISBN: 0465029973.

[26] A. L. Samuel, "Programming Computers to Play Games," *Advances in Computers*, vol. 1, no. C, pp. 165–192, 1960, ISSN: 00652458. DOI: 10.1016/S0065-2458(08)60608-7.

[27] E. H. Shortliffe, "A rule-based computer program for advising physicians regarding antimicrobial therapy selection," in *Proceedings of the 1974 Annual ACM Conference, ACM 1974*, Association for Computing Machinery, Inc, 1974, p. 739. DOI: 10.1145/1408800.1408906.

[28] R. Davis and J. J. King, "The Origin of Rule-Based Systems in AI," Tech. Rep., 2005.

[29] M. A. Shwe, B Middleton, D. E. Heckerman, M Henrion, E. J. Horvitz, H. P. Lehmann, and G. F. Cooper, "Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. The probabilistic model and inference algorithms.," *Methods of information in medicine*, vol. 30, no. 4, pp. 241–55, 1991, ISSN: 0026-1270. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1762578.

[30] J. S. Brown, *Pedagogical, natural language, and knowledge engineering techniques in SOPHIE-I, II and III*, 1982.

[31] G. Montavon, W. Samek, and K. R. Müller, *Methods for interpreting and understanding deep neural networks*, 2018. DOI: 10.1016/j.dsp.2017.10.011. arXiv: 1706.07979.

[32] W. J. Clancey, "From Guidon To Neomycin and Heracles in Twenty Short Lessons: Orn Final Report 1979-1985.," *AI Magazine*, vol. 7, no. 3, pp. 40–60, 187, 1986, ISSN: 07384602.

[33] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein Macrocognition, "DARPA XAI Literature Review p. Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI Prepared by Task Area 2 Institute for Human and Machine Cognition," Tech. Rep., 2019.

[34] M. S. Mahoney, "The History of Computing in the History of Technology," *IEEE Annals of the History of Computing*, vol. 10, no. 2, 1988.

[35] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations : An Approach to Evaluating Interpretability of Machine Learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2008. arXiv: arXiv:1806.00069v2.

[36] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018, ISSN: 21693536. DOI: 10.1109/ACCESS.2018.2870052.

[37] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," no. October, 2019. arXiv: 1910.10045. [Online]. Available: http://arxiv.org/abs/1910.10045.

[38] T. Kulesza, M. Burnett, W. K. Wong, and S. Stumpf, "Principles of Explanatory Debugging to personalize interactive machine learning," in *International Conference on Intelligent User Interfaces, Proceedings IUI*, vol. 2015-Janua, Association for Computing Machinery, 2015, pp. 126–137, ISBN: 9781450333061. DOI: 10.1145/2678025.2701399.

[39] D. Slack, S. A. Friedler, C. Scheidegger, and C. D. Roy, "Assessing the Local Interpretability of Machine Learning Models," Tech. Rep., 2019. arXiv: 1902.03501v2.

[40] A. Kirsch, "Explain to whom? Putting the user in the center of explainable AI," *CEUR Workshop Proceedings*, vol. 2071, 2018, ISSN: 16130073.

[41] M. Ribera and A. Lapedriza, "Can we do better explanations? A proposal of User-Centered Explainable AI," Tech. Rep., 2019.

[42] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing Design Practices for Explainable AI User Experiences," 2020. DOI: 10.1145/3313831.3376590. arXiv: 2001.02478v1.

[43] C. Yang, A. Rangarajan, and S. Ranka, "Visual Explanations From Deep 3D Convolutional Neural Networks for Alzheimer's Disease Classification," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2018, pp. 1571–1580, 2018, ISSN: 1942597X. arXiv: 1803.02544.

[44] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Confounding variables can degrade generalization performance of radiological deep learning models," pp. 1–15, 2018. DOI: 10.1371/journal.pmed.1002683. arXiv: 1807.00431. [Online]. Available: http://arxiv.org/abs/1807.00431{\%}0Ahttp://dx.doi.org/10.1371/journal.pmed.1002683.

[45] A. W. Thomas, H. R. Heekeren, K. R. Müller, and W. Samek, "Analyzing Neuroimaging Data Through Recurrent Deep Learning Models," *Frontiers in Neuroscience*, vol. 13, no. Dl, pp. 1–36, 2019, ISSN: 1662453X. DOI: 10.3389/fnins.2019.01321. arXiv: 1810.09945.

[46] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, pp. 1–8, 2014. arXiv: 1312.6034.

[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020, ISSN: 15731405. DOI: 10.1007/s11263-019-01228-7. arXiv: 1610.02391.

[48] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K. R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, pp. 1–8, 2019, ISSN: 20411723. DOI: 10.1038/s41467-019-08987-4. arXiv: 1902.10178. [Online]. Available: http://dx.doi.org/10.1038/s41467-019-08987-4.

[49] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, "Interpretable Deep Models for ICU Outcome Prediction," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2016, pp. 371–380, 2016, ISSN: 1942597X.

[50] P. Rafi, A. Pakbin, and S. Kumar Pentyala, "Interpretable Deep Learning Framework for Predicting all-cause 30-day ICU Readmissions," Tech. Rep., 2017.

[51] S. M. Lauritsen, M. Kristensen, M. V. Olsen, M. S. Larsen, K. M. Lauritsen, M. J. Jørgensen, J. Lange, and B. Thiesson, "Explainable Artificial Intelligence Model to Predict Acute Critical Illness from Ekectronic Health Records," Tech. Rep., 2019. arXiv: 1912.01266v1.

[52] E. Zihni, V. I. Madai, M. Livne, I. Galinovic, A. A. Khalil, J. B. Fiebach, and D. Frey, "Opening the Black Box of Artificial Intelligence for Clinical Decision Support: A Study Predicting Stroke Outcome," 2020. DOI: 10.1101/19010975. [Online]. Available: http://dx.doi.org/10.1101/19010975.

[53] J. B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," *Artificial Intelligence in Medicine*, 2019, ISSN: 18732860. DOI: 10.1016/j.artmed.2019.01.001.

[54] M. R. Zafar and N. M. Khan, "DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems," 2019. arXiv: 1906.10263. [Online]. Available: http://arxiv.org/abs/1906.10263.

[55] Y. Belinkov and J. Glass, "Analysis Methods in Neural Language Processing: A Survey," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 49–72, 2019, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00254. arXiv: 1812.08951.

[56] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 1101–1111, 2018. DOI: 10.18653/v1/n18-1100. arXiv: 1802.05695.

[57] A. H. Gee, D. Garcia-Olano, J. Ghosh, and D. Paydarfar, "Explaining deep classification of time-series data with learned prototypes," *CEUR Workshop Proceedings*, vol. 2429, pp. 15–22, 2019, ISSN: 16130073. arXiv: 1904.08935.

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

[59] L. Lin, B. Xu, W. Wu, T. Richardson, and E. A. Bernal, "Medical Time Series Classification with Hierarchical Attention-based Temporal Convolutional Networks: A Case Study of Myotonic Dystrophy Diagnosis," 2019. arXiv: 1903.11748. [Online]. Available: http://arxiv.org/abs/1903.11748.

[60] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 3530–3537, 2018. arXiv: 1710.04806.

[61] L. Arras, F. Horn, G. Montavon, K. R. Müller, and W. Samek, ""WHAT IS RELEVANT IN A TEXT DOCUMENT?": AN INTERPRETABLE MACHINE LEARNING APPROACH," *PLoS ONE*, vol. 12, no. 8, pp. 1–19, 2017, ISSN: 19326203. DOI: 10.1371/journal.pone.0181142. arXiv: 1612.07843.

[62] F. Horst, S. Lapuschkin, W. Samek, K. R. Müller, and W. I. Schöllhorn, "Explaining the unique nature of individual gait patterns with deep learning," *Scientific Reports*, vol. 9, no. 1, 2019, ISSN: 20452322. DOI: 10.1038/s41598-019-38748-8. arXiv: 1808.04308.

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

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

[65] M. Guilleme, V. Masson, L. Roze, and A. Termier, "Agnostic local explanation for time series classification," *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, vol. 2019-November, pp. 432–439, 2019, ISSN: 10823409. DOI: 10.1109/ICTAI.2019.00067.

[66] L. Ye and 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*, ser. KDD '09, Paris, France: Association for Computing Machinery, 2009, 947–956, ISBN: 9781605584959. DOI: 10.1145/1557019.1557122. [Online]. Available: https://doi.org/10.1145/1557019.1557122.

[67] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim, "Towards a rigorous evaluation of XAI methods on time series," *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, no. Ml, pp. 4197–4201, 2019. DOI: 10.1109/ICCVW.2019.00516. arXiv: 1909.07082.

[68] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K. W. Low, S. F. Newman, J. Kim, and S. I. Lee, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 749–760, 2018, ISSN: 2157846X. DOI: 10.1038/s41551-018-0304-0. [Online]. Available: http://dx.doi.org/10.1038/s41551-018-0304-0.

[69] D. Price, "How to read an electrocardiogram (ECG). Part 1: Basic principles of the ECG. The normal ECG," *Southern Sudan Medical Journal*, vol. 3, no. 2, 2010.

[70] U. Rajendra Acharya, J. S. Suri, J. A. Spaan, and S. M. Krishnan, *Advances in cardiac signal processing*, U. R. Acharya, J. S. Suri, J. A. E. Spaan, and S. M. Krishnan, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, , pp. 55–81, ISBN: 3540366741. DOI: 10.1007/978-3-540-36675-1. [Online]. Available: http://link.springer.com/10.1007/978-3-540-36675-1.

[71] K. Casey, *How to explain machine learning in plain English | The Enterprisers Project*. [Online]. Available: https://enterprisersproject.com/article/2019/7/machine-learning-explained-plain-english (visited on 01/15/2020).

[72] C. Molnar, *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*. 2019, https://christophm.github.io/interpretable-ml-book/.

[73] M. E. Morocho-cayamcela, H. Lee, and W. Lim, "Machine Learning for 5G / B5G Mobile and Wireless Communications : Potential , Limitations , and Future Directions," *IEEE Access*, vol. 7, no. September, pp. 137 184–137 206, 2019. DOI: 10.1109/ACCESS.2019.2942390.

[74] J. Brownlee, *Basic Concepts in Machine Learning*. [Online]. Available: https://machinelearningmastery.com/basic-concepts-in-machine-learning/ (visited on 01/16/2020).

[75] S. Asari, *Machine Learning Classifiers - Towards Data Science*. [Online]. Available: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623 (visited on 01/16/2020).

[76] M. Silva, *Interpreting Machine Learning Model - Towards Data Science*. [Online]. Available: https://towardsdatascience.com/interpreting-machine-learning-model-70fa49d20af1 (visited on 01/16/2020).

[77] S. Singh, *Why correlation does not imply causation? - Towards Data Science*. [Online]. Available: https://towardsdatascience.com/why-correlation-does-not-imply-causation-5b99790df07e (visited on 01/23/2020).

[78] C. E. Brodley and M. A. Friedl, "Decision tree classification of land cover from remotely sensed data," *Remote Sensing of Environment*, vol. 61, no. 3, pp. 399–409, 1997, ISSN: 00344257. DOI: 10.1016/S0034-4257(97)00049-7.

[79] J. Tamibmaniam, N. Hussin, W. K. Cheah, K. S. Ng, and P. Muninathan, "Proposal of a Clinical Decision Tree Algorithm Using Factors Associated with Severe Dengue Infection," *PLOS ONE*, vol. 11, no. 8, S. D. Sekaran, Ed., e0161696, 2016, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0161696. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0161696.

[80] H. Kato, H. Hanada, and I. Takeuchi, "Learning sparse optimal rule fit by safe screening," *ArXiv*, vol. abs/1810.01683, 2018.

[81] O. M'Haimdat, *Understand the Fundamentals of the K-Nearest Neighbors (KNN) Algorithm*. [Online]. Available: https://heartbeat.fritz.ai/understand-the-fundamentals-of-the-k-nearest-neighbors-knn-algorithm-533dc0c2f45a (visited on 11/23/2020).

[82] P. Cunningham and S. J. Delany, *K-nearest neighbour classifiers: 2nd edition (with python examples)*, 2020. arXiv: 2004.04523 [cs.LG].

[83] J. Grabocka, N. Schilling, M. Wistuba, and L. Schmidt-Thieme, "Learning time-series shapelets," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14, New York, New York, USA: Association for Computing Machinery, 2014, 392–401, ISBN: 9781450329569. DOI: 10.1145/2623330.2623613. [Online]. Available: https://doi.org/10.1145/2623330.2623613.

[84] C. Walker, *White Box vs Black Box Models: Balancing Interpretability and Accuracy*, 2020. [Online]. Available: https://blog.dataiku.com/white-box-vs-black-box-models-balancing-interpretability-and-accuracy (visited on 09/09/2020).

[85] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?," 2017. arXiv: 1712.09923. [Online]. Available: http://arxiv.org/abs/1712.09923.

[86] N. Hatami, Y. Gavet, and J. Debayle, "Classification of time-series images using deep convolutional neural networks," in *Tenth International Conference on Machine Vision (ICMV 2017)*, A. Verikas, P. Radeva, D. Nikolaev, and J. Zhou, Eds., International Society for Optics and Photonics, vol. 10696, SPIE, 2018, pp. 242 –249. DOI: 10.1117/12.2309486. [Online]. Available: https://doi.org/10.1117/12.2309486.

[87] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification," *Journal of Systems Engineering and Electronics*, vol. 28, no. 1, pp. 162–169, 2017.

[88] B. Rocca, *Handling Imbalanced Datasets in Machine Learning*, 2019. [Online]. Available: https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28 (visited on 08/10/2020).

[89] DARPA, "Broad Agency Announcement: Explainable Artificial Intelligence (XAI)," pp. 1–52, 2016, ISSN: 15580644. DOI: 10.1109/36.210458. [Online]. Available: https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf{\%}0Ahttp://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf{\%}0Apapers3://publication/uuid/AF93BD83-DA5C-48BE-A0B4-212DC3C78A31.

[90] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing Theory-Driven User-Centric Explainable AI," *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pp. 1–15, 2019. DOI: 10.1145/3290605.3300831. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3290605.3300831.

[91] D. Alvarez-Melis and T. S. Jaakkola, *On the robustness of interpretability methods*, 2018. arXiv: 1806.08049 [cs.LG].

[92] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-Agnostic Interpretability of Machine Learning," no. Whi, 2016. arXiv: 1606.05386. [Online]. Available: http://arxiv.org/abs/1606.05386.

[93] Bleckwen, *Interpretability Of Machine Learning Models – Part 2*, 2019. [Online]. Available: https://bleckwen.ai/interpretability-of-machine-learning-models-part-2/ (visited on 09/16/2020).

[94] L. Breiman, "Random forests," *Machine Learning*, vol. 1, no. 45, pp. 5–32, 2001. DOI: 10.1201/9780367816377-11.

[95] 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, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[96] M. T. Ribeiro, S. Singh, and 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, ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939778.

[97] G. J. Katuwal and R. Chen, "Machine Learning Model Interpretability for Precision Medicine," 2016. arXiv: 1610.09045. [Online]. Available: http://arxiv.org/abs/1610.09045.

[98] J. Borovec, J. Švihlík, J. Kybic, and D. Habart, "Supervised and unsupervised segmentation using superpixels, model estimation, and graph cut," *Journal of Electronic Imaging*, vol. 26, no. 06, p. 1, 2017, ISSN: 1017-9909. DOI: 10.1117/1.jei.26.6.061610.

[99] P. Kopper, "Lime and neighbourhood," in *Limitations of Interpretable Machine Learning Methods*, C. Molnar, Ed., 2019, ch. 13, pp. 201–222. [Online]. Available: https://compstat-lmu.github.io/iml{_}methods{_}limitations/.

[100] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, and H. Gamboa, "Tsfel: Time series feature extraction library," *SoftwareX*, vol. 11, p. 100 456, 2020. DOI: 10.1016/j.softx.2020.100456.

[101] J. Rodrigues, D. Folgado, D. Belo, and H. Gamboa, "SSTS: A syntactic tool for pattern search on time series," *Information Processing and Management*, vol. 56, no. 1, pp. 61–76, 2019, ISSN: 03064573. DOI: 10.1016/j.ipm.2018.09.001. [Online]. Available: https://doi.org/10.1016/j.ipm.2018.09.001.

[102] J. Sarkar and C. Peterson, "Enabling Prognostics of Robust Design with Interpretable Machine Learning," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2019-December, pp. 286–289, 2019, ISSN: 01631918. DOI: 10.1109/IEDM19573.2019.8993481.

[103] C. Molnar, G. König, B. Bischl, and G. Casalicchio, "Model-agnostic Feature Importance and Effects with Dependent Features – A Conditional Subgroup Approach," pp. 1–20, 2020. arXiv: 2006.04628. [Online]. Available: http://arxiv.org/abs/2006.04628.

[104] C. Schockaert, P. W. S. A, A. Schmitz, and P. W. S. A, "VAE-LIME : Deep Generative Model Based Approach for Local Data-Driven Model Interpretability Applied to the Ironmaking Industry,"

[105] L. Hu, J. Chen, V. N. Nair, and A. Sudjianto, "Locally Interpretable Models and Effects based on Supervised Partitioning (LIME-SUP)," pp. 1–15, 2018. arXiv: 1806.00663. [Online]. Available: http://arxiv.org/abs/1806.00663.

[106] R. Elshawi, Y. Sherif, M. Al-Mallah, and S. Sakr, "Interpretability in healthcare a comparative study of local machine learning interpretability techniques," *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, vol. 2019-June, pp. 275–280, 2019, ISSN: 10637125. DOI: 10.1109/CBMS.2019.00065.

[107] Y. R. Xie, D. C. Castro, S. E. Bell, S. S. Rubakhin, and J. V. Sweedler, "Single-Cell Classification Using Mass Spectrometry through Interpretable Machine Learning," *Analytical Chemistry*, 2020, ISSN: 0003-2700. DOI: 10.1021/acs.analchem.0c01660.

[108] T. Górecki and M. Łuczak, "First and second derivatives in time series classification using dtw," *Communications in Statistics-Simulation and Computation*, vol. 43, no. 9, pp. 2081–2092, 2014.

[109] T. Gorecki and M. Luczak, "Using derivatives in time series classification," *Data Mining and Knowledge Discovery*, vol. 26, no. 2, pp. 310–331, 2013.

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

[111] D. Folgado, M. Barandas, R. Matias, R. Martins, M. Carvalho, and H. Gamboa, "Time alignment measurement for time series," *Pattern Recognition*, vol. 81, pp. 268–279, 2018.

[112] M. Thomas, M. K. Das, and S. Ari, "Automatic ECG arrhythmia classification using dual tree complex wavelet based features," *AEU - International Journal of Electronics and Communications*, vol. 69, no. 4, pp. 715–721, 2015, ISSN: 16180399. DOI: 10.1016/j.aeue.2014.12.013.

[113] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG heartbeat classification: A deep transferable representation," *CoRR*, vol. abs/1805.00794, 2018. arXiv: 1805.00794. [Online]. Available: http://arxiv.org/abs/1805.00794.

[114] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and 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*, 92–100, 2019, ISSN: 1611-3349. DOI: 10.1007/978-3-030-32245-8_11. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-32245-8_11.

[115] Y. Yuan, M. Chao, and Y. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE Transactions on Medical Imaging*, vol. 36, no. 9, pp. 1876–1886, 2017.

[116] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods, "Tslearn, a machine learning toolkit for time series data," *Journal of Machine Learning Research*, vol. 21, no. 118, pp. 1–6, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-091.html.

[117] "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals.," *Circulation*, vol. 101, no. 23, 2000, ISSN: 15244539. DOI: 10.1161/01.cir.101.23.e215.

[118] A. for the Advancement of Medical Instrumentation and A. N. S. Institute, *Testing and Reporting Performance Results of Cardiac Rhythm and ST-segment Measurement Algorithms*, ser. ANSI/AAMI. The Association, 1998, ISBN: 9781570201165.

[119] P. De Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 7, pp. 1196–1206, 2004, ISSN: 00189294. DOI: 10.1109/TBME.2004.827359. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/15248536/.

[120] V. Mondéjar-Guerra, J. Novo, J. Rouco, M. G. Penedo, and M. Ortega, "Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers," *Biomedical Signal Processing and Control*, vol. 47, pp. 41–48, 2019, ISSN: 17468108. DOI: 10.1016/j.bspc.2018.08.007.