Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier xxx.DOI

Explainable AI for Time Series Classification: A review, taxonomy and research directions

ANDREAS THEISSLER[1](Member, IEEE), FRANCESCO SPINNATO[2], UDO SCHLEGEL[3], AND RICCARDO GUIDOTTI[4]

[1]Aalen University of Applied Sciences (e-mail: andreas.theissler@hs-aalen.de)
[2]Scuola Normale Superiore (e-mail: francesco.spinnato@sns.it)
[3]University of Konstanz (e-mail: u.schlegel@uni-konstanz.de)
[4]University of Pisa (e-mail: riccardo.guidotti@unipi.it)

Corresponding author: Andreas Theissler (e-mail: andreas.theissler@hs-aalen.de).

ABSTRACT Time series data is increasingly used in a wide range of fields, and it is often relied on in crucial applications and high-stakes decision-making. For instance, sensors generate time series data to recognize different types of anomalies through automatic decision-making systems. Typically, these systems are realized with machine learning models that achieve top-tier performance on time series classification tasks. Unfortunately, the logic behind their prediction is opaque and hard to understand from a human standpoint. Recently, we observed a consistent increase in the development of explanation methods for time series classification justifying the need to structure and review the field. In this work, we (a) present the first extensive literature review on Explainable AI (XAI) for time series classification, (b) categorize the research field through a taxonomy subdividing the methods into time points-based, subsequences-based and instance-based, and (c) identify open research directions regarding the type of explanations and the evaluation of explanations and interpretability.

INDEX TERMS Explainable Artificial Intelligence, Time Series Classification, Interpretable Machine Learning, Temporal Data Analysis

I. INTRODUCTION

MACHINE LEARNING (ML) models have achieved unprecedented performance in recent years. While the models become more accurate and complex, the lack of model explainability or interpretability is one of the key challenges of ML research. Such a challenge may prevent the use of ML in applications that call for interpretable decisions, such as high-stakes fields like healthcare or autonomous systems [1]. For this reason, and due to the state-of-the-art performance of these models in many other areas, there is a need to overcome this problem. The research field of eXplainable AI (XAI) [2], [3] or interpretable machine learning [4] tackles explainability challenges to give insights into model behavior.

A large part of the work in explainability is done on tabular data or in the field of computer vision, where deep neural networks (DNN) typically achieve state-of-the-art performance. While computer vision is undeniably an important research field of machine learning, we argue that there might be a bias in XAI research toward image data due to *(i)* the availability of data, e.g. Imagenet [5] or CIFAR-10 [6] and – more importantly – *(ii)* the inherent semantics present in images: explaining the classification of a rooster based on the rooster comb is easily interpretable and verifiable, while a time series is often not intelligible without domain knowledge [7].

We believe that time series should receive the same research attention since they are omnipresent, e.g., in technical systems [8], [9], the medical domain [10], or business applications [11]. Further, due to the tremendous amount of data generated by sensors over time, machine learning models yield superior results at many tasks due to their capability to capture long- as well as short-term patterns in the data [12]. Thus, such models can outperform experts in certain time series tasks, enabling their application in various use cases, e.g., in predictive maintenance [8], [13], heartbeat anomaly detection [10], or texture recognition [14]. The research field of XAI for time series classification has become more popular since around 2019, a variety of valuable papers have been published in recent years (see Figure 1). This trend was the

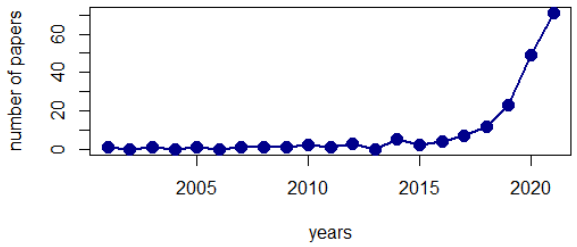


FIGURE 1. The number of papers published per year on XAI for time series classification started to increase significantly in 2019, suggesting an increase in the topic's relevance. The search was performed on Scopus using the search terms presented in Table 1.

motivation to structure the field with a review of the most important works and to deduce open research directions to close gaps.

The primary goals of this work are to *(i)* give an overview of the current body of literature on XAI for Time Series Classification (TSC), *(ii)* categorize the research field through a sound taxonomy, and *(iii)* deduce new insights, identifying open research challenges in order to inspire new research in this emerging field. We achieve these goals by surveying papers in the field. Thus, we conduct a semi-structured literature review by including works we consider influential and a systematic search with Scopus to systematically create an overview of the field.

We further introduce applications and evaluations which incorporate the reviewed approaches or can be applied to them. Specifically, we present a set of applications where the aforementioned methods have been used in various areas to include explainability in applications. We discuss the state-of-the-art regarding the evaluation of the aforementioned XAI approaches and also include references to applicable, but not yet implemented, computer vision evaluation techniques. Applications and evaluations give a valuable insight into the state-of-the-art of the discussed XAI approaches and introduce further opportunities on how to deploy these techniques.

Lastly, we discuss the findings of our selected papers and propose future opportunities for XAI for TSC. The final discussion presents the trends and challenges we identified in our review of the different approaches towards explainability for time series models. Furthermore, based on these challenges and other influences, we highlight future research directions to contribute closing the gaps we identified. Thus, in the following, we contribute:

1) a *semi-structured literature review* of the most recent explainable AI approaches for time series classification;
2) a *taxonomy* of approaches for XAI deduced from the reviewed work
3) insights into the differences and advantages of such explainable AI techniques;
4) highlights of applications and evaluation strategies to showcase applied XAI techniques;
5) research directions in order to inspire future research in the field of XAI for time series classification.

The rest of the paper is organized as follows. Section II illustrates recent surveys in the areas of XAI, time series, and works at their intersection. Section III reports basic notions and definitions necessary to understand the content of this survey. Section IV details the research methodology adopted to retrieve the works presented in this survey. Section V presents the proposed taxonomy and the review, while Section VI presents applications of the XAI methods reviewed and discusses evaluation of explanations. Finally, Section VII discusses findings and illustrates future research directions in XAI for TSC.

II. RELATED WORK

The intensive request for explainability approaches [15] largely contributed to the massive increase of research in XAI. The proliferation of XAI methods working in different domains has been accompanied by various surveys categorizing these methodologies [2], [3], [16], [17]. An introduction to frequently used explainers in XAI can for example be found in the books [18], [19] and in the surveys [2], [3]. However, while explainers for data types such as relational data, images, and texts are illustrated from various perspectives in different literature reviews, explainers for other data types, like time series, are not reviewed sufficiently in detail. In the rest of this section, we report general surveys on XAI not specifically addressing time series, surveys on TSC, and two preprints of surveys on explainability methods for TSC, highlighting the differences to our paper.

In [2], a classification of XAI methods according to the problem they are able to solve is presented. The first categorization is between (1) *explanation by design* or *intrinsic interpretability*, and (2) *black-box explanation* or *post-hoc explanation*. In [3], [16], [17] the same principal categorization is adopted. The second categorization further classifies the black-box explanation problem into model explanation, outcome explanation, and black-box inspection.

Another significant distinction shared among [2], [3], [17], [20], [21] is between *model-specific* and *model-agnostic* explanation methods. In this survey, we adopt and exploit the same taxonomy of [2], [3], [17], [20], [21] which is detailed in the next section. However, while these surveys are generalists, we focus on explainers for time series classification problems. We underline that some surveys related to XAI are focused not only on machine learning but also on social studies [22], [23], recommendation systems [24], model-agents [25], and domain-specific applications such as health and medicine [26] or predictive maintenance [27].

Concerning surveys for TSC not addressing XAI, the works of [28], [29] and [30] are probably most updated and complete. In [28] the focus is more on classical approaches, implementing and comparing 18 algorithms starting from the simple and popular k-Nearest Neighbor (kNN) and then illustrating more novel and complex classifiers. On the other hand, in [29], the focus is on neural network-based approaches, and the performance of deep learning algorithms are presented with an empirical study involving the most recent deep neural

network (DNN) architectures for TSC. A detailed analysis of time series classifiers based on Convolutional Neural Networks (CNNs) is presented in [31]. In [30] the focus is on multivariate time series classification, comparing 16 state-of-the-art TSC algorithms. However, none of the surveys above touches on questions related to interpretability or explainability. The authors of [32] presented a focused review on time series classifiers adopting a distance-based approach, as well as a discussion of the strengths and weaknesses of each method and distance measure reviewed. Distance-based classifiers can be considered transparent if it is possible to retrieve the most similar time series responsible for the classification and if the distance measure is simple enough. However, most advanced distance-based approaches use complex distance measures like Dynamic Time Warping [33], and therefore they are omitted from this survey.

To the best of our knowledge, the only existing review papers at the intersection of XAI and TSC are the pre-prints [34], [35]. The authors of [34] present an overview of XAI methods for TSC and illustrate the types of explanations they produce. In their overview table, they categorize XAI methods by the type of model to be explained, i.e., CNNs or RNNs (Recurrent Neural Networks), whereas we focus on the type of explanation returned by the explainers. In addition, differently from [34], we also discuss evaluation measures for explainers of time series classifiers. In [35], XAI with respect to TSC is faced at a high level and the survey only reports *i)* generalist explanation methods such as LIME [36], SHAP [37], Grad-CAM [38] and DeepLIFT [39], *ii)* explanation methods for neural networks. In contrast to [35], we focus more on explanation methods designed explicitly for TSC, including many different kinds of XAI approaches, such as transparent models and non-neural network-based methods, thus providing an extended overview of the state-of-the-art.

III. SETTING THE STAGE

This section introduces notations and definitions useful to comprehend the state-of-the-art. First, we report definitions for time series classification, and then we formalize the concepts related to explainability.

A. TIME SERIES CLASSIFICATION

This section presents formal definitions for Time Series Classification (TSC) and recalls basic notions. We define a time series as follows:

Definition 1. A *time series* $x = \{t_1, t_2, \ldots, t_m\} \in \mathbb{R}^{m \times d}$ is an ordered set of m real-valued observations (or time steps), with dimensionality d.

We say that a time series is *univariate* when $d = 1$, i.e., each observation $t_i \in \mathbb{R}$ is a real value. On the other hand, when $d > 1$ we name x a *multivariate time series* (also referred to as *multidimensional time series*), i.e., each observation $t_i \in \mathbb{R}^d$ is a vector containing multiple real values. From another perspective, a multivariate time series is formed by d univariate time series with length m. Often, the univariate time series which are part of a multivariate time series are also referred to as *signals*, or *channels* [40].

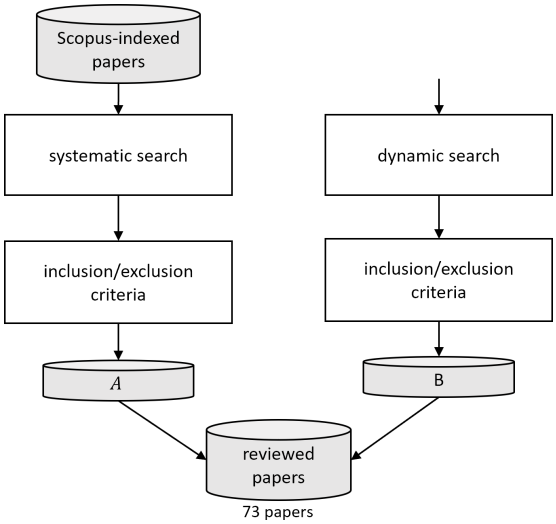


FIGURE 2. Review methodology: Semi-systematic literature review combining a systematic search on Scopus (Table 1) with a dynamic search. Exclusion and inclusion criteria (Table 2) were applied to identify the final set of papers.

A set of time series, either univariate or multivariate, with attached labels, forms a time series classification dataset.

Definition 2. A *time series classification dataset* $D = (X, Y)$ is a set of n time series, $X = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^{n \times m \times d}$, with a vector of assigned labels (or classes), $Y = \{y_1, y_2, \ldots, y_n\} \in \mathbb{N}^n$.

For a dataset D containing l classes, y_i can take l different values. When $l = 2$, D is a binary classification dataset, while for $l > 2$, D is a multi-class classification dataset. We can now define the TSC problem as:

Definition 3. Given a TSC dataset D, *Time Series Classification* is the task of training a function or mapping f from the space of possible inputs X to a probability distribution over the class values Y.

The resulting TSC function f takes as input a time series x and returns the label y of the class to which x belongs to according to what f learned, i.e., $y = f(x)$. We use $f(X) = Y$ as a shorthand for $\{f(x) \mid x \in X\} = Y$. Typically, the classifier f can be queried at will.

B. EXPLAINABLE ARTIFICIAL INTELLIGENCE

The research field of Explainable AI (XAI) studies approaches that unveil the logic behind automatic decision-making systems [2], [41]–[43]. In general, we like to point out that the terminology in XAI is not fully established yet (see, e.g., the study of [44]). Some researchers use some terms interchangeably, while others view them as different. An example are the terms *explainability* and *interpretability*. While we do not aim to establish a new terminology, we give some insight into the current state of the discussion. One definition states that XAI has the goal of creating "a suite of

new or modified ML techniques that produce explainable models that, when combined with effective explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems" [45]. In [46], XAI is described as a tool "to ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms.". Regarding the two common terms *explainability* and *interpretability*, we follow the definition of [47] stating that "systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation". The definition in [2] is in line with that above, stating that interpretability describes the extent to which a model and/or its predictions are human-understandable. Models can be categorized into those that provide interpretability themselves (sometimes referred to as white-box models or intrinsically interpretable models) and those requiring an explanation (commonly referred to as black-box models). This distinction is made in many papers [1], [2], [22]. Interpretability can be viewed as a *passive* characteristic and explainability as an *active* characteristic of a model or method [48].

An interesting new perspective is brought up in [49], where the term *quasi-explanations* is introduced. This refers to explanations that include terms foreign to the domain for which the explanation is intended. In other words, the target user will not be able to understand these quasi-explanations. The authors of [49] state that even models frequently referred to as intrinsically interpretable models might not be interpretable for the end user. Eventually, one line of research advocates not using black-box models for critical decisions but rather building intrinsically interpretable models [1]. In this paper, we do not aim to take a stand for or against using black-box models. However, we believe that when using black-box models for important decisions, explanations should be provided that are interpretable in the sense outlined above.

According to the current XAI literature, XAI approaches can be categorized according to different criteria. In our survey, we follow the categorization presented in [2] and [3]:

Ante-hoc vs. Post-hoc: Ante-hoc explainable methods, such as decision trees, are models that can be considered directly interpretable due to their simple structure and/or transparency by design. However, we like to note that transparency in that sense has been discussed at three levels in [50]. Adding explainability to a black-box does not necessarily make the model interpretable as a whole but rather sheds light on specific parts of the model or the model's decisions.

Post-hoc explainability approaches are instead separated from the model they explain and can provide insight into what a model has learned after training without changing its underlying structure, e.g., LIME [36]. Defining an intrinsic explainable method means learning a classification function f that directly unveils the reasons for the classification. On the other hand, a post-hoc explainability method should be applied when f is a black-box model like an artificial neural network (ANN), a support vector machine (SVM), or a random forest, and the reason for the decision is not directly accessible or understandable. Thus, a post-hoc explainer typically consists of a function g that takes as input the classifier f as well as a dataset D. We highlight that in our review, we survey both ante-hoc [1] and post-hoc [2] approaches that have been proposed or can be utilized for time series classification.

TABLE 1. Systematic search on Scopus (title, abstract, keywords). The rows were combined with AND operators into one search query.

criterion	search terms
XAI	(interpretab* OR explainab* OR XAI)
ML	("machine learning" OR "deep learning" OR "artificial intelligence" OR "AI" OR "neural network")
TSC	(classif*) AND ("time series")
type	journal article OR conference paper
language	English

Global vs. Local: Global explanation approaches provide an explanation that describes the overall logic of the entire model for any input instance, i.e., g returns a generalized explanation for the decisions that are valid for the whole set X. On the other hand, local explanation approaches explain the behavior of a model for a specific instance, i.e., g unveils the reasons for the classification only for a specific instance x.

Model-Agnostic vs. Model-Specific: Model-agnostic explainers g can be used to explain any type of classifier f, i.e., it does not matter if f is an ANN, a Random Forest, or a composition of private software for decision-making. LIME [36] is a well-known example of a model-agnostic explainer. Model-specific explainers g are specifically built to add interpretability to a certain type of classifier f, i.e., g are only able to explain a classifier f belonging to a specific family of classifiers. For instance, Grad-CAM [38] is able to explain only differentiable classifiers like CNNs. We claim that every ante-hoc approach is, by definition, model-specific, given that it can be used only to explain itself, i.e. $f = g$.

IV. REVIEW METHODOLOGY

The overriding goals of our paper are to *(i)* give an overview, *(ii)* categorize, and *(iii)* deduce new insights from the current body of literature on XAI for time series classification. These goals are achieved by reviewing papers in the field.

While a systematic literature search might seem like a natural choice, we found that it will yield an incomplete review in this emerging field. Reasons are different terminologies used in different research subfields. Examples are the papers on shapelets that are quite different from deep learning papers. Hence, we opted for a semi-systematic literature review [51]:

1) we conducted a systematic search on Scopus using a set of search terms (Table 1);
2) we conducted a dynamic search to uncover additional papers in the different subfields;
3) the found papers were judged by the authors based on exclusion and inclusion criteria (Table 2) in order to decide whether to include a paper.

TABLE 2. List of inclusion and exclusion criteria, where IC refer to inclusion and EC to exclusion criteria, respectively.

ID	criterion
IC1	- only time series classification, no forecasting
IC2	- anomaly detection papers only if achieved by some sort of classification with supervised learning
IC3	- only papers that explicitly address and enhance explainability or interpretability
IC4	- only work on raw time series data including the time-frequency domain, no hand-crafted features
IC5	- only papers that show their approach for time series or are trivially adaptable
EC1	- no preprints
EC2	- no papers published prior to 2011
EC3	- no papers without any citations
EC4	- no surveys or reviews (these would be included in the related work section, though)
EC5	- no papers that conduct explorative analyses on the statistics of inner network components
EC6	- no papers massively relying on domain- or application-specific characteristics of the time series data
EC7	- no papers on time series streams, online or real-time classification

The review methodology is shown in Figure 2.

In order for the reader to be able to distinguish papers included in the review from papers supplying background information, we reference reviewed papers with author names. For example, Gee et al. [52] references a reviewed paper while [2] is referenced for background information. Furthermore, all reviewed papers are shown in Table 3.

V. XAI FOR TIME SERIES CLASSIFICATION

In the following, we highlight the advantages and characteristics of our semi-systematic literature review. We report the papers analyzed in Table 3. The papers are organized according to the following *taxonomy* (see Figure 3). First, we discriminate on the granularity of explanation returned depending on the portion of a time series used to illustrate the causes for the decision process. Accordingly, we have three *families* of explanation methods characterized by the type of explanation returned (Figure 3). In particular, we have recognized *time points-based explanations* if the explanation refers to specific time points in a time series, *subsequences-based explanations* if the explanation refers to sub-parts of the time series, and *instance-based explanations* if the explanation adopts entire time series as explanation. We put into the *others* family those explanation types that cannot be tied to any of the previous ones. We increase the detail of the taxonomy by further analyzing and categorizing the XAI approaches falling into the three aforementioned families. Indeed, for each family of XAI methods, we further differentiate the algorithmic strategies adopted by the reviewed approaches to return the explanation. Details are provided in the respective subsections. In addition, in Table 3 we further categorize the explanation methods into ante-hoc and post-hoc approaches, into model-specific and model-agnostic approaches, and into global and local explanations. Additional categories include whether the approaches were explicitly designed for time series (TS-specific or not) and whether they were designed for univariate or multivariate TS.

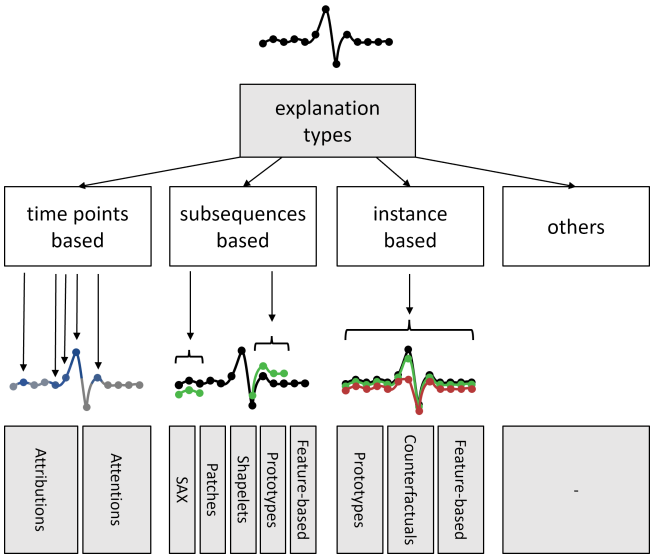


FIGURE 3. The proposed taxonomy categorizes the reviewed XAI approaches in different explanation types based on their explanations. The explanation methods used to generate the explanations are assigned into our categories.

Also, for every method, we report the name (if available), the reference, the publication year, and the code language with a hyperlink to the corresponding library.

A. TIME POINTS-BASED EXPLANATIONS

Explanations based on *time points* assign a relevance score or weight to every time point of a time series. Such scores indicate how much a certain time point contributed to the model's decision. Formally, we define time points-based explanations as:

Definition 4 (Time Points-based Explanation)**.** Given a time series x, a *time points-based explanation* $e = \{r_{i,j} \mid \forall i \in [1, m], j \in [1, d]\}$ for the time series x contains a relevance score $r_{i,j}$ for every real-valued input data point $t_{i,j}$ of x where the index i refers to the time point and j to the dimension in case of a multivariate time series.

Such relevance scores can be retrieved in different ways. The most widely adopted ones are related to approaches based on *attributions* and *attentions*. At a high level, we can say that attribution-based approaches exploit some *external* method, which uses the TSC model to attribute output predictions to input variables. On the other hand, attention-based approaches use some *internal* mechanism of the TSC model to show which variables of the input they use. In both cases, interpretability is achieved by considering the most important time points and presenting these to users, e.g., visually.

Attributions. Attributions methods are often deployed in computer vision, as one can visualize the output as a heatmap to gain insights into the model's relevant regions in the input [120]. However, even attribution techniques typically used in computer vision, such as LIME [36] or LRP [55],

TABLE 3. List of reviewed papers and taxonomy for XAI methods for Time Series Classification. Table legend: Post/Ante-hoc: P-Post, A-Ante; Model-Agnostic/Specific: A-Agnostic, S-Specific; Global/Local: G-Global, L-Local; TS-Specific: (✓) if it is a Time Series specific method; Uni/Multivariate: U-Univariate, M-Multivariate; Code: P-Python, M-Matlab, J-Java, JS-Javascript. Code letters are hyperlinks to official web pages. We considered all Ante-hoc methods as Model-Specific methods.

Name	Ref	Year	Explanation Type	Explanation Method	Post/Ante -hoc	Model- Agnostic/Specific	Global/ Local	TS-Specific	Uni/Multi -variate	Code (URL)
Integrated Gradients	[53]	2017	Time Points-based	Attributions	P	S	L	✗	-	P
FCN	[54]	2017		Attributions	P	S	L	✓	U	P
LIME	[36]	2016		Attributions	P	A	L	✗	-	P
LRP	[55]	2015		Attributions	P	S	L	✗	-	P
ExcitationBP	[56]	2016		Attributions	P	S	L	✗	-	P
Occlusion	[57]	2014		Attributions	P	S	L	✗	-	P
SHAP	[37]	2017		Attributions	P	A	L	✗	-	P
SmoothGrad	[58]	2017		Attributions	P	S	L	✗	-	P
DeepLIFT	[39]	2019		Attributions	P	S	L	✗	-	P
Salience-CAM	[59]	2021		Attributions	P	S	L	✓	U	-
Grad-CAM	[60]	2020		Attributions	P	S	L	✗	-	P
TSViz	[61]	2019		Attributions	P	S	L	✓	U	P
TSXplain	[62]	2019		Attributions	P	S	L	✓	M	-
TSInsight	[63]	2021		Attributions	P	S	L	✓	M	-
SoundLIME	[64]	2017		Attributions	P	A	L	✓	U	P
MTEX-CNN	[65]	2019		Attributions	P	S	L	✓	M	-
PERT	[66]	2021		Attributions	P	A	L	✓	U	P
FIT	[67]	2020		Attributions	P	S	L	✓	M	P
WinIT	[68]	2021		Attributions	P	S	L	✓	M	P
CEFEs	[69]	2021		Attributions	P	S	L	✓	U	-
XTF-CNN	[70]	2021		Attributions	P	S	L	✓	U	-
LEFTIST	[71]	2019		Attributions	P	A	L	✓	U	P
ALSTM-FCN	[12]	2017		Attentions	P	S	L	✓	U	P
GCRNN	[72]	2018		Attentions	A	S	L	✓	U	P
-	[73]	2018		Attentions	P	S	L	✓	U	P
ETSCM	[74]	2019		Attentions	A	S	L	✓	U	-
DACNN	[75]	2020		Attentions	A	S	L	✓	M	-
LAXCAT	[76]	2021		Attentions	A	S	G	✓	M	-
DeepVix	[77]	2020		Attentions	A	S	G	✓	M	JS
VixLSTM	[78]	2021		Attentions	P	S	G	✓	M	-
-	[79]	2021		Attentions	A	S	L	✓	M	P
-	[10]	2014	Subsequences-based	SAX	A	S	G	✓	U	-
SAX-VSM	[80]	2013		SAX	A	S	G	✓	U	-
-	[81]	2020		SAX	A	S	G	✓	M	P
MR-SEQL	[82]	2020		SAX	A	S	G	✓	U	P
CPHAP	[83]	2021		Shapelets	P	S	L	✓	U	-
Shapelets	[84]	2011		Shapelets	A	S	G	✓	U	J
ShapeletTransform	[85]	2012		Shapelets	A	S	G	✓	U	-
MSD	[86]	2012		Shapelets	A	S	G	✓	M	-
LS	[87]	2011		Shapelets	A	S	G	✓	U	-
LTS	[88]	2014		Shapelets	A	S	G	✓	U	P
LCTS	[89]	2016		Shapelets	A	S	G	✓	U	J
-	[90]	2018		Shapelets	A	S	G	✓	U	C++
-	[91]	2018		Shapelets	A	S	G	✓	U	-
LRS	[92]	2019		Shapelets	A	S	G	✓	U	P
ADSNs	[93]	2020		Shapelets	A	S	G	✓	U	P
XCNN	[94]	2020		Shapelets	A	S	L	✓	U	-
GENDIS	[95]	2021		Shapelets	A	S	G	✓	U	P
GMSM	[96]	2021		Shapelets	A	S	G	✓	M	-
MAPIC	[97]	2021		Shapelets	A	S	G	✓	U	P
DASH	[98]	2021		Shapelets	A	S	G	✓	U	P
TORRENT	[99]	2021		Shapelets	P	S	G	✓	U	-
LASTS	[100]	2020		Shapelets	P	A	L	✓	U	P
PatchX	[101]	2021	Instance-based	Patches	A	S	L	✓	M	-
P2ExNet	[102]	2020		Prototypes	A	S	L	✓	M	-
ProtoFac	[103]	2020		Prototypes	P	S	G	✗	-	P
mWDN	[104]	2018		Feature-based	A	S	G	✓	U	P
ProSeNet	[105]	2019		Prototypes	A	S	G	✓	U	-
-	[52]	2019		Prototypes	A	S	G	✗	U	JS
DPSN	[106]	2020		Prototypes	P	S	G	✓	U	P
TapNet	[107]	2020		Prototypes	P	S	G	✓	M	P
MODL-TSC	[108]	2013		Feature-based	A	S	G	✓	U	-
MTDT	[109]	2018		Feature-based	A	S	G	✓	U	R
FC/TAA/LTAA	[110]	2019		Feature-based	A	S	G	✓	M	-
Conceptual	[111]	2020		Feature-based	P	A	G	✓	U	-
-	[112]	2021		Feature-based	A	S	G	✓	U	-
Native Guide	[113]	2021		Counterfactuals	P	A	L	✓	U	P
τ_{RT}/τ_{IRT}	[114]	2020		Counterfactuals	P	S	L	✓	U	P
CoMTE	[115]	2021		Counterfactuals	P	A	L	✓	M	P
CEM	[116]	2020		Counterfactuals	P	S	L	✗	-	-
RCN	[117]	2019	Others	Rules	A	S	G	✓	U	-
TCCL	[118]	2020		Granger Causality	A	S	G	✓	M	-
-	[119]	2020		Temporal Logic	A	S	G	✓	U	-

can be applied to time series to better understand the model's behavior [7], [120]. Attributions techniques can be categorized into three classes, *gradient-based*, *structure-based*, and *surrogate-and-sampling-based* [120]. Gradient-based methods (Integrated Gradients [53] proposed by Sundararajan *et al.*, Grad-CAM by Selvaraju *et al.* [60] adapted for time series in [54] by Wang *et al.*, SmoothGrad by Smilkov *et al.* [58], Saliency [121]) use the gradients of the input with regard to the output to get attributions. While structure-based techniques (LRP by Bach *et al.* [55], DeepLIFT by Shrikumar *et al.* [39], Excitation Backpropagation by Zhang *et al.* [56]) use a score which gets backpropagated from the output to the input. Finally, surrogate-and-sampling methods (Ribeiro *et al.*'s LIME [36], Lundberg *et al.*'s SHAP [37] and Occlusion by Zeiler *et al.* [57]) generate samples around the given input to train an interpretable model or use a game-theoretical weighting of the features to gain attributions. A more sophisticated approach as opposed to using data points is, similarly to superpixels for image data, to first segment the time series and then use each segment of the time series as a feature [71], [100].

The authors of [7] propose to use current *computer vision* techniques such as the previously introduced attribution techniques like LIME, SHAP and others to explain deep learning models for time series in the same way as for images. They evaluated LIME, LRP, DeepLIFT, Saliency, and SHAP against each other as well as a random explanation. They further propose evaluation techniques using a perturbation analysis on the produced attributions. The time points within subsequences of the input data were perturbed for a selection of ten univariate time series datasets. A decrease in the accuracy caused by the perturbation is assumed to indicate that an important part of the time series for the models' prediction power was altered. The results are reported for three ML models: a CNN, an RNN, and ResNet. While there is no clear best XAI method for the CNN or RNN in terms of the decrease induced by the perturbation, for the more advanced ResNet, SHAP shows the highest average decrease across all data sets and perturbation settings. Hence, of the evaluated XAI methods, SHAP appears to best capture the time points relevant for classification behavior of the model.

Also, in [59], computer vision techniques (CAM and Grad-CAM) are used to explain time series classifiers. In particular, Zhou *et al.* enhance CAM and Grad-CAM with a backpropagation to combine saliency as well as CAM calculations and improve the generalizability of CAM. They show on six datasets how the approach *Salience-CAM* outperforms CAM with improved attributions.

In [61], Siddiqui *et al.* propose *TSViz* explaining CNNs by showing which regions in the input data are responsible for a decision (saliency maps) as well as the influence of the network's filters on a given decision. The explanations are based on the layers' gradients, i.e., they use a gradient$\times$input [121] approach. In [62], Munir *et al.* exploit TSViz to design *TSXplain*, a system for time series explanation of DNN decisions. TSXplain finds the most salient regions responsible for a certain prediction and the most important time series through TSViz [61]. Such regions and instances are then combined with different statistical features used to generate natural language explanations. In a user study, the explanations were provided to expert and novice users, and the majority of users were satisfied with the explanations. The textual explanations differ from most other explainers, and we view them as promising. The authors acknowledge that their explanation system is task-specific and cannot easily be transferred to a different task. A further evolution of TSViz is *TSInsight* [63] a post-hoc explainer for TSC proposed by Siddiqui *et al.* TSInsight trains an autoencoder (AE) on the input data. The AE is fine-tuned using the trained classifier's gradients, and the AE's objective function is enhanced by a sparsity-inducing norm driving the AE to reproduce the relevant parts of the input time series.

In [64], Mishra *et al.* propose *SoundLIME*, a perturbation-based method that, using LIME, explains TSC in the field of music content analysis. Explanations are based on patterns, where, as a key contribution in addition to the temporal domain, also the frequency and the time-frequency domain are incorporated into the explanation. Assaf *et al.* propose *MTEX-CNN* [65], an end-to-end explainable CNN that can classify multivariate time series and simultaneously generate saliency maps. MTEX-CNN uses a two-stage network architecture combined with specific kernel sizes, allowing the application of Grad-CAM for visualizing the attention over both time and dimensions. In [66], Parvatharaju *et al.* proposed a perturbation-based method named *PERT* to find interesting and relevant time points. For a given time series, PERT first finds time series in the data that can be used as background to perform perturbations. Then, it learns the extent to which each time step can be perturbed without altering the prediction of the classifiers. The goal of PERT is to keep the number of perturbed data points minimal. In [67], Tonekaboni *et al.* propose Feature Importance in Time (*FIT*) to observe the temporal shift influence of individual features over time to estimate their importance. FIT contrasts the predictive distribution of a model against a counterfactual, using the contribution of the predictive distributional shift under a KL-divergence. Thus, FIT observes the model's behavior under the influence of fixing some variables and changing others. In [68], Rooke *et al.* extend FIT by proposing *WinIT* that directly uses the predictive distributional shift on the explanation with lookback-windows to further enable longer-lasting patterns. A perturbation analysis shows that WinIT improves the explanations of FIT on a medical dataset they selected, while experiments on a synthetic dataset show that the explanations returned are correct.

While we exclude from the survey papers strongly relying on application-specific features, in the following, we briefly illustrate the contribution of [69], [70] that exploit domain-knowledge, not in the feature extraction step, but to evaluate the explanations. In [69], a 1D-CNN is trained on univariate ECG time series, and the data is transformed into the time-frequency domain while Grad-CAM is used to explain the

classification. Then, the difference between the explanation and the well-known features used by clinicians is quantified. This quantification is used to validate the learned representations and explanations generated by the 1D-CNN. In [70], a dual-channel 1D-CNN is used to detect rock fracturing in univariate time series: one channel is used to process the temporal domain, and the other to process the frequency domain. The explanation is based on Grad-CAM [60] and is evaluated by visualized examples and, in addition, w.r.t. to domain knowledge. While these approaches are not directly generalizable, we believe the ideas are transferable to selected domains, where the underlying data is well-understood and shows clear, acknowledged patterns.

In [71], Guilleme *et al.* propose *LEFTIST* a model-agnostic local explainer. LEFTIST segments the input time series uniformly and creates neighborhoods by replacing some of these interpretable components with a transform function, e.g., by constant values or random background from training data. This way, the importance of the segments of the time series is determined. The explanation is then obtained from these components using LIME or SHAP. This method is somewhat between a time point-based and a subsequences-based one. Indeed, the final explanation is in the form of the importance of segments of the time series, but can also be viewed as a saliency map highlighting time points inside each segment. We highlight that this paper includes an extensive study on interpretability: fidelity analysis is conducted, comparing the approach to a white-box and a more complex black-box model. Furthermore, the interpretability is assessed as helpful in a user study with 194 participants.

Summary and analysis. Attributions are used to attribute a relevance score to each input value of a model. In many fields, such as computer vision, these attribution techniques are straightforward to implement and fast to compute. Furthermore, these methods can extract regions of an input that are important for understanding the focus of the model. In time series, such relevance scores can also be calculated with the same computational time as for images [7]. However, due to the non-intelligible nature of time series, generating explanations is much more difficult using just attributions and their relevance scores [120]. When applied to time series, attributions and their primary explanation medium, heatmaps, are often promising for domain experts but ineffective for general users, given that the relevance scores are difficult to interpret without additional knowledge about the underlying data [122]. Regarding surrogate-and-sampling methods that can be applied to time series, considering each time step as a feature, we like to point the reader to two recent papers. [123] discusses challenges of LIME, SHAP, and related methods, independent of their use on time series. They emphasize the known fact that the methods' underlying assumption is feature independence. However, feature independence is not respected for adjacent observations in a time series. The authors of [124] stress that Shapley values, which are the fundamentals of SHAP, assume that adding players to the game does not decrease its overall value. However, adding features to an ML model may decrease the model's performance.

Attentions. In contrast to attributions, attentions use an internal mechanism to incorporate a special focus, i.e., attention, of the network onto certain parts of the input or transformed data. Long Short-term Memory (LSTM) encoder-decoder architectures (Seq2Seq-Models) calibrate important areas of the input for the decoder to involve the overall context [125]. Later, attention layers were introduced, focusing even heavier on the internal calibration of the input data towards itself (attention and self-attention) [126]. These attention mechanisms are often used in transformer networks to achieve state-of-the-art performance in language tasks [126]. However, attention should be handled with care, as it sometimes does not directly show the relevant parts of the input for the classification and can be attacked easily with adversarial examples [127]. Attention approaches are designed to work with different types of deep learning architectures, as shown in the following. Most of the approaches in this section are ante-hoc explanation methods, i.e., the attention is embedded in the network architecture.

Karim *et al.* [12] combine CNNs with LSTM submodules to create a specialized time series classification model. While the main contribution of the work is the classifier itself, the authors propose a variant incorporating an attention mechanism, allowing to explain the decision process of the LSTM cell. In [72], Lin *et al.* propose *GCRNN*, a Group-Constrained Convolutional Recurrent Neural Network. GCRNN comprises three modules: CNN, RNN, and SGL, the latter being a fully connected module with a group lasso penalty. The CNN module extracts high-level features while the RNN module learns the temporal characteristics of the data. Finally, the purpose of the SGL is to reduce the complexity of the model by regularizing it while also allowing the inspection of its attention regions. Interpretability is assessed based on a model-internal metric; however, its evaluation might be future work. In [73], Vinayavekhin *et al.* propose a temporal contextual layer that incorporates an attention mechanism into time series classification. In contrast to using recurrent layers, they propose to provide the whole input sequence to the attention layer. This layer calculates each attention weight based on information of the input time-series. This allows the network to directly select dependencies in the data and to assign significant weight only to the most important time steps. The model is evaluated with three brief case studies on sequential data but not on the type of time series we refer to in our paper. In [74], Hsu *et al.* propose *ETSCM* (Explainable Time Series Classification Model) that can perform interpretable early classification of multivariate time series. ETSCM first employs a pre-trained deep learning method to extract the features among the different time series dimensions and capture their temporal structure. Then, an attention mechanism highlights the most important segments for the classification. The authors were able to conduct an expert-led study. For ECG time series, two medical doctors evaluated the interpretability of the produced explanations. The doctors claimed that the

highlighted sections had no medical significance because no entire intervals were highlighted. In a second dataset, the doctors confirmed that approximately half of the provided explanations were "correct" from a medical standpoint.

In [75], Hosseini *et al.* propose *DACNN*, a deep-aligned CNN specifically aimed at tackling SBARS, i.e., skeleton-based action recognition and segmentation. *DACNN* contains so-called alignment filters that can extract and highlight important local patterns in the temporal dimensions of the data more efficiently than regular convolution filters. Even if the paper shows an application-specific approach, these filters could be utilized in a more general setting to increase the interpretability of CNNs. In [76], Hsieh *et al.* introduce *LAXCAT*, a Locality Aware eXplainable Convolutional ATtention network that is able to classify multivariate time series transparently. The framework is composed of a CNN feature extraction module and two attention modules that can identify the key variables responsible for the classification and the most discriminative time intervals. In [77], Dang *et al.* propose *DeepVix*, an LSTM model that supports interactive operations allowing a visual representation of the intermediate steps of the learning process. DeepVix enables the user to perform what-if analyses on multivariate time series to understand the most important features and customize the neural network configurations by injecting domain knowledge. DeepVix was extended in [78] (*VixLSTM*) to incorporate Shapley Values, improving the usability of the framework. Finally, Schwenke *et al.* [79] turn time series into a symbolic form and then train a transformer model. The data points are subdivided based on their attention score and are either fully included, partially included, or discarded using two user-defined thresholds. Regarding evaluation, the authors took an interesting approach: the method is evaluated by training a classification model on the transformed data and comparing its performance with a model trained on the original data. The underlying assumption is that if the results do not change significantly, the transformer does indeed focus on the relevant parts. Note that while this approach segments the time series in a first step, due to the used explanation method, it was categorized as time points-based rather than subsequences-based.

Summary and analysis. Similar to attributions, attentions can show relevant parts of the input. However, attention only works if specific components are implemented into the model's architecture. Also, like attributions, attentions are often visualized as heatmaps and are somewhat hard to interpret in many cases [122]. Further, these explanations can be misleading, as different attentions (meaningful and not meaningful) can still produce the same output [127]. An interesting insight comes from [74], where domain experts (medical doctors) were not satisfied with the highlighted subset of the data points because the selection did not correspond with units typically considered by domain experts.

Summary of time points-based methods. The reviewed time points-based XAI methods comprise attribution methods incorporating an attention mechanism, with about two-thirds of the papers presenting attribution methods. While some methods were initially not developed for time series, e.g., SHAP and LIME, the research community has developed specific methods for time series. The proposed methods address univariate and multivariate time series in equal shares.

B. SUBSEQUENCES-BASED EXPLANATIONS

Explanations based on subsequences identify sub-parts of a time series responsible for the classification outcomes. Formally, we define a subsequence as:

Definition 5 (Subsequence)**.** Given a time series $x = \{t_1, \ldots, t_m\}$, a *subsequence* $s = \{t'_i, \ldots, t'_{i+l-1}\}$ of length l is an ordered sequence of values such that $1 \leq i \leq m - l + 1$.

We further distinguish subsequences as *proper* and *improper* as follows. A *proper subsequence* of the time series x is a direct and continuous sampling of values from x, i.e. $s = \{t_i, \ldots, t_{i+l-1}\}$. In other words, the set of observations in a proper subsequence is part of a time series. On the other hand, an *improper subsequence* of the time series x is a subsequence s for which there is no requirement of correspondence between the observation of the subsequence and the observations in the time series. Usually, an improper subsequence holds some semantic meaning with respect to a time series dataset, identifying, for example, important patterns that are similar in different time series with small shape changes. These kinds of patterns are sometimes also called subsequence prototypes, not to be confused with time series prototypes that are defined in Section V-C.

Furthermore, subsequences can be *real-valued* or *discretized*. Real-valued subsequences are directly extracted from the raw time series. On the other hand, discrete subsequences are commonly obtained through the Symbolic Aggregate approXimation (SAX) algorithm [128]. In short, SAX transforms time series into strings. The algorithm uses Piecewise Aggregate Approximation (PAA) [129] to discretize the time series, dividing it into equally sized bins and averaging the values of each bin. Then, the PAA segments are converted into a sequence of symbols, usually letters. This approximation can reduce noise and capture the main characteristics of the time series. Discretized subsequences are, for the most part, proper subsequences because their symbolic representation can be mapped back to the original segments in the time series. Both for proper and improper subsequences, interpretability is achieved by considering the most discriminative sequences of time points.

Proper Subsequences. TSC methods relying on proper subsequences (sometimes also named patches or patterns) are typically interpretable-by-design approaches.

Due to possible computational problems in extracting subsequences of the most suitable length, often approximations such as SAX are performed, as detailed in the following. In [10], Maletzke *et al.* propose to extract *motifs*, i.e. repeating patterns, and *characteristics*, i.e. global statistics,

from the times series and to train interpretable symbolic models on them. The approach identifies both local and global patterns by extracting all subsequences of a given length from the time series, compressing them using *SAX*. Then, a Collision Matrix is built to identify subsequences likely to be motifs. These patterns are then used with decision trees to achieve interpretable classification. In [80], Senin *et al.* extract characteristic patterns using *SAX* with an overlapping window. From SAX patterns, a bag-of-patterns model is created. Then, a class is described by discriminative patterns extracted from all the time series of this class. New time series are classified w.r.t. the discriminative patterns per class. Song *et al.* [81] proposed an approach for multivariate times series that learns a representation based on deconvolutions and bag-of-words that were created from *SAX* subsequences. Classification is conducted with logistic regression on the learned representation vectors and the bag-of-words. The approach is evaluated with classification accuracy and with author-selected examples visualized as a network plot. The interpretability of this plot is, however, not evaluated.

Similarly to [80], Nguyen *et al.* [82] represent the time series in a symbolic form with multiple resolutions. In addition, Symbolic Fourier Approximation (SFA) [130] is used to incorporate the frequency domain. New time series are then classified based on the found representation using a linear classifier. While the paper's main focus is the improvement of the classification accuracy, interpretability is shown for a variant of the approach, excluding the difficult to interpret frequency domain induced by the SFA approach.

The approach of Cho *et al.* in [83] extracts subsequences by first considering highly activated nodes in a CNN, followed by the clustering of the extracted subsequences with a self-organizing map (SOM) approach. The subsequences identified are those more responsible for the activations. Therefore, this approach uses an attention approach, but the explanation of the classification is returned in terms of subsequences for the various layers. In [101], Mercier *et al.* transform a time series sample into patches by splitting the initial sample. These patches are used to train a DNN to classify the overall sample. Using the highest predicted class of the DNN for the individual patches and connecting them, they train another ML model on these extracted predictions to classify the overall sample. The framework returns the most important patches and, while being interpretable, is only slightly worse than a simple DNN directly trained on the original data.

Summary and analysis. Proper subsequences are parts of the original time series extracted during training. They highlight time series intervals that are important for the model and can assist in combining domain knowledge with model knowledge extraction on data. At first sight, they should have a high interpretable power due to their interpretability by design, directly pointing users to relevant parts of the original data. However, such subsequences, patches, or patterns are often still relatively hard to interpret due to the non-intelligibility of time series. For this reason, these approaches need further tweaks to perform better, and abstractions need to be easier to interpret for users. For example, motif-based models do not have state-of-the-art accuracy with small, easy-to-understand decision trees [10]. A challenge with using subsequences is the segmentation of the data. Often, fixed-length intervals are used, and the choice of the length is critical since, for a suboptimal choice, segmentation may not capture relevant patterns.

Improper Subsequences – Shapelets. Shapelets were first introduced in [131] as a new time series primitive embedded in a decision tree classifier. Shapelets are sequences of values that are most representative of class membership, i.e., depending on their distance from the time series, they split the dataset, maximizing the information gain. At each step of the decision tree induction, both the shapelet and the best dataset split point are determined. Interpretable classification is then achieved by inspecting the decision tree and the shapelets in each tree node. Formally, we can define shapelets as:

Definition 6 (Univariate Shapelet)**.** Given a TSC dataset $D \in \mathbb{R}^{n \times m \times d}$, a *univariate shapelet* $s \in \mathbb{R}^l$ of length $l < m$ is an improper subsequence that discriminates the target Y.

Typically, a shapelet-based method extracts a set containing k-most discriminative shapelets, denoted as $S \in \mathbb{R}^{k \times l \times d}$, typically with $k < n$ and $d = 1$ for univariate time series. The interpretable classifiers, such as decision trees but also logistic regressors or others, are then trained on the so-called shapelet transformation approach [85]:

Definition 7 (Shapelet Transformation)**.** Given a set of time series X, a set of shapelets S, and a distance function $dist$, a *shapelet transformation* turns a set of time series $X \in \mathbb{R}^{n \times m \times d}$ into a matrix of continuous values $X' \in \mathbb{R}^{n \times k}$, obtained by taking the minimum distance $dist$ between each time series in X, and each shapelet in S.

The shapelet transform extracts the k most discriminative shapelets from a time series dataset and returns a new representation of the data, where the attributes represent the distances between the k shapelets and each time series. In this way, any classification algorithm can be used, potentially increasing the accuracy while reducing training time. In general, the choice of the distance measure is method-dependent. In the case of multivariate time series and univariate shapelets, it is calculated w.r.t. the dimension the shapelet was extracted from [40]. In [86], the problem of shapelet extraction in a multivariate setting is first tackled. The proposed method is called Multivariate Shapelets Detection (MSD) and extracts shapelets from all time series dimensions. The approach uses an information gain-based distance to split the dataset and ranks the shapelets depending on a utility score, weighted to favor shapelets appearing earlier.

Definition 8 (Multivariate Shapelet)**.** Given a TSC dataset $D \in \mathbb{R}^{n \times m \times d}$, a *multivariate shapelet* $s \in \mathbb{R}^{l \times d}$ is a set of d aligned univariate shapelets of equal length $l < m$.

Since the introduction of shapelets in [131], there were

many contributions focused on improving the efficiency of the shapelet search both for univariate [132]–[139], and multivariate [140]–[142] time series. In the following, we include only the approaches that contribute from an interpretability standpoint and not necessarily from an efficiency standpoint.

In [87], Mueen *et al.* propose *Logical-Shapelets (LS)*, a more expressive classification approach that addresses the problem of scalability of the original method. Furthermore, LS can exploit conjunctions or disjunctions of shapelets to discriminate the target variable. In [88], Grabocka *et al.* formalize the shapelet search as an optimization problem that jointly learns the shapelets from the training data and minimizes their incurred error without the need to explore all possible candidates. The approach, called *Learning-Shapelets (LTS)*, first roughly estimates the shapelets and then iteratively learns and optimizes their shape via gradient descent by minimizing a classification loss function. LTS was extended in [143] to use Dynamic Time Warping as a distance measure. In [89], Yang *et al.* propose *LCTS*, a shapelet learning method that, instead of extracting the top shapelets directly from time series subsequences, uses self-organizing incremental neural networks (SOINN) to first generate shapelet prototypes. The learned candidates are then used to transform the time series into the shapelet feature space by combining an exponential function with distance normalization. Finally, an L1-regularizer is used to select the top shapelets from the candidates. The advantage of using SOINN is that isolated shapelets are removed and similar shapelets combined, resulting in a smaller and better-quality set of shapelets. We believe that a smaller but more precise set of shapelets leads to better interpretability since it reduces the cognitive workload of end users.

In [90], Fang *et al.* propose an efficient way to learn shapelets using a multi-stage process on the PAA-transformed and the raw time series. Interpretability is explicitly addressed by finding shapelets so that each shapelet represents one characteristic of a class. This property is achieved by incorporating a coverage metric for the class characteristics in a final filtering process on the previously found shapelets. Classification is performed with a linear classifier. In [91] another prototype-based shapelet learning approach is proposed by Deng *et al.* This method is based on dictionary learning theory and learns basic representative shapes robust to deformations and transformations. It first minimizes the average least-squares error between the transformed subsequences and the shapelet prototypes. It then updates the dictionaries to represent the basic shapes learned from the time series. The discovered shapelet prototypes are reported to be more general and expressive because they preserve the intrinsic shapes present in the data. In [92], Guilleme *et al.* propose the method *Localized Random Shapelets*. This approach aims to generate more realistic and interpretable shapelets by adding shapelet localization to the traditional shapelet transform representation. Using a hierarchical feature selection process with regularization, the approach can be tuned to select, for each shapelet, either only its distance information or both distance and localization information. In this way, the user can understand how much the localization of the shapelet, besides its presence in the time series, is important for the prediction.

In [93], Ma *et al.* propose *Adversarial Dynamic Shapelet Networks (ADSNs)*. This approach dynamically generates shapelets that are more similar to real subsequences by framing the shapelet generation process as a two-player minimax game, following the idea of Generative Adversarial Networks (GAN). The discriminator is trained to distinguish between synthetic shapelets and real subsequences in the input time series. A regularization term is added to the objective function to avoid model collapse and ensure shapelet diversity. Analogously to prototypes, we believe that diverse shapelets lead to better interpretability. In [94], Wang *et al.* propose *XCNN*, an adversarily regularized *EXplainable Convolutional Neural Network*. XCNN learns discriminative and meaningful shapelets by using two networks: a CNN to classify the time series and a discriminator to regularize the classifier and force it to learn shapelets similar to real subsequences of the training set. In [95], Vandewiele *et al.* propose *GENDIS*, a genetic approach that uses evolutionary computation to perform the shapelet search. One of the key advantages of this method is that it evaluates entire sets of shapelets instead of independently analyzing single shapelet candidates. In this way, both the quality of the candidate sets and their size can be optimized, resulting in fewer and more different shapelets. Moreover, GENDIS allows taking into account interactions between shapelets explicitly. In [96], Medico *et al.* present a DNN in which multivariate shapelets are embedded as trainable weights. By adding two regularization terms to the loss function, the approach can efficiently and transparently classify multivariate time series, retrieving a small set of uncorrelated shapelets.

In [97], Guidotti and D'Onofrio show an alternative approach to designing an interpretable shapelet-based decision tree. In particular, the proposed approach (*MAPIC*) exploits the Matrix Profile [144] to extract shapelets in the form of motifs and discords for each splitting of the decision tree. The approach is efficient both theoretically and empirically, while being comparable or even outperforming approaches using different procedures for shapelet extraction.

In [99], Hu *et al.* propose a combination of shapelets and attention-mechanism. An efficient shapelet transformation aims to reduce the number of shapelets. While the interpretability of the shapelets is not explicitly addressed, the reduction of the number of shapelets is in line with the concept of sparsity, known from prototype methods. Classification is conducted based on the found shapelets combined with an RNN with an attention mechanism. The use of the RNN improves the classification accuracy yet, turns the approach into a model-specific approach.

In [100] Guidotti *et al.* proposed *LASTS*, a Local Agnostic Shapelet-based Time Series explainer. LASTS uses an autoencoder first to compress the time series into a simplified latent encoding. Then, a genetic algorithm generates synthetic

instances that, once decoded, constitute a neighborhood containing both prototypical and counterfactual time series. Finally, a shapelet-based decision tree is trained to output factual and counterfactual rules, explaining the classification in terms of subsequences that must, or must not, be contained to get a specific black-box outcome.

Summary and analysis. Shapelets are a compromise between state-of-the-art performance (e.g., accuracy) and interpretability. In some cases, the first learning shapelet approaches did not explicitly address interpretability, such as [142]. The accuracy of shapelet methods has increased for later methods like [93] due to various extensions. However, the pipeline as a whole has become more difficult to understand as a result of the inclusion of complex models, such as an autoencoder [100] or a GAN [93]. Furthermore, being closer to the data, such approaches push shapelets closer to proper subsequences by addressing issues such as the fact that they are still difficult to interpret for some use cases and users. Discovering meaningful shapelets in multivariate time series is particularly challenging due to potentially differing change points in the time series of the different dimensions.

Improper Subsequences – Prototypes. Prototypes are improper archetypal subsequences for building interpretable time series classifiers. Ming *et al.* proposed *ProSeNet* [105] which encodes sequential data, not constrained to time series, with an encoding network. The encoded representation is then passed to a layer that learns prototypes fulfilling the criteria of simplicity, diversity, and sparsity. A novelty compared to previous work is using a constrained similarity measure, rather than the commonly used L2 distance, to compare the encodings with the learned prototypes. Mercier *et al.* [102] train an autoencoder to generate embeddings for an input time series in their approach *P2ExNet*. The embedding representation is then fed into a prototype network in which multiple subsequence prototypes of the whole input time series, instead of a single prototype, are used. Following that, a softmax-layer conducts the classification based on the prototypes. The results show only a marginal performance decrease for the accuracy w.r.t. models not using prototypes.

Summary and analysis. One idea of prototype-based explanations is to combine domain knowledge with learning approaches such as neural networks to generate meaningful domain-specific subsequences. Interestingly, all methods reviewed in this subsection extract prototypes by means of neural networks. Prototype explanations generally incorporate knowledge about experts' data acquired during previous analysis or through know-how. Furthermore, black-box models incorporating expert-created prototypes are sometimes applied [105]. In summary, prototypes propose a promising extension for opaque models but need other techniques to open the black-boxes around them.

Improper Subsequences – Feature-based. In [104], Wang *et al.* propose to use multilevel discrete wavelet transform to decompose time series into subsequences ranked by the contained frequencies. The data is then classified in the time-frequency domain by a cascade of classifiers (different types of neural networks) incorporating residual connections. In doing so, the time-frequency domain is included at different resolutions. Classification is explained by the importance of each time-frequency part.

Summary and analysis. Feature-based methods on improper subsequences extract features from some kind of transformed representation rather than the original data points. For example, the frequency-domain was used in [104]. Regarding explainability, the features' interpretability determines the explanation's quality. Features from the frequency domain could be used as an example in fields where frequency bands are part of the domain-specific knowledge, such as EEG frequency bands in the medical domain.

Summary of subsequences-based methods. The types of subsequences-based methods are manifold, comprising methods based on the SAX representation as well as shapelets. The reviewed papers addressed patches, prototypes, and feature-based methods to a lesser extent. Except for one, all methods were originally designed for time series data. Furthermore, we highlight that the great majority of the interpretable subsequences-based classifiers make use of shapelets. Indeed, shapelets are a powerful discriminative tool that can be theoretically and practically extended to other data types. In turn, this allows experimenting with these interpretable features on domains different from time series [98]. Although shapelets could be considered state-of-the-art in this category, this might indicate a research gap in developing other types of interpretable subsequences-based time series classifiers.

C. INSTANCE-BASED EXPLANATIONS

The methods falling in this category rely on the whole time series instance to express the reasons for the classification. We mainly recognize two categories of methods returning instance-based explanations. On one hand, those counting on features extracted from the whole time series. On the other hand, those returning time series instances as explanations. Regarding the latter, the most common explanations are prototypes and counterfactuals.

Feature-based Explanations. Feature-based interpretability approaches try to explain classification, e.g., through statistics extracted from the time series data. These features are not based on individual time points or subsequences and therefore are usually less sensitive to noise.

An early representative for the *extraction of feature sets* is the work of Gay *et al.* [108]. Time series are transformed into multiple representations using generic transformations like derivatives or auto-correlation functions to identify patterns characteristic for each class. Co-clustering is employed to group similar patterns within classes and thereby identify a set of class-discriminant features. Classification is then conducted on a feature space with standard classifiers, shown with Naive Bayes. The approach is applied to univariate time series, but it can be transferred to multivariate time series by adapting the feature extraction steps. Interpretability is achieved by con-

sidering the identified features in a textual form. Evaluating the interpretability of the created explanations could be future work.

In [109], Shalaeva *et al.* propose *MTDT* (Multi-operator Temporal Decision Trees) extending the decision tree algorithm to time series data. *MTDT* uses split operators to capture different geometrical structures in the data based on dynamic time warping and spherical operators besides SAX subsequences. Users can then inspect each node condition in the decision tree to understand how the classification was achieved.

In [110], Ito and Chakraborty focus on computational efficiency and present three shape-aware feature extraction methods with linear time-complexity, that compute the similarities between time series. The Fold Count (FC) representation counts the number of foldings of a time series on itself. The Time Axes Area (TAA) relaxes FC by measuring the areas under the folds. The Log Weighted Area is a modification of TAA that takes the logarithm to avoid too long delays. As an interpretable classifier, kNN is adopted to exploit these shape-aware time series representations.

In [111], Küsters *et al.* introduce a framework to extract, test, and evaluate the intrinsic features used by models *adopting pre-defined filters*. Filters, such as lowpass filters, can be applied to the input data before the classifier's prediction to investigate the model's behavior regarding these intrinsic features. Such variables can be used to compare experts' domain knowledge with the features used by the model. The approach is evaluated against LRP [55] and shows that by changing fewer data in a fidelity analysis, the accuracy changes even more.

Extracting statistical features, the work of Zaman *et al.* [112] builds a decision tree for the classification of control chart patterns (univariate time series). The decision tree is assumed to be interpretable and is shown as an example.

Summary and analysis. Feature-based explanations have a long history and are traditionally applied to time series classification. However, in many cases, such approaches are hard to understand for non-expert users, and in some cases, even for experts. For example, as observed for MTDT [109], the resulting decision trees are still hard to explore. Due to their intrinsic use of features in the time series, these feature-based approaches hold promising value for evaluating other explanations, such as attributions, like in Küsters *et al.* [111]. In summary, the interpretability of feature-based methods strongly depends on the interpretability of the used features for the target users.

Prototype-based Explanations. Prototypical examples are time series exemplifying the main aspects responsible for a classifier's specific decision outcome. Formally,

Definition 9 (Prototype)**.** Given a classifier f, an instance $\tilde{x}$ is a prototype if there is a set of instances $X' \subset X$ *represented by* $\tilde{x}$, and such that $\forall x \in X', f(\tilde{x}) = f(x)$.

A prototype can be a real record sampled from the dataset that is important and meaningful because it summarizes the shape of many other similar instances, or a synthetic one, for example a cluster centroid or a record generated by following some ad-hoc processes. The explanation is obtained by comparing an instance x for which we have the decision $f(x)$ with the prototype $\tilde{x}$.

In [52], Gee *et al.* propose an approach for learning time series prototypes. The *prototypes are found using an autoencoder*, and the work's novelty is the learning of diverse prototypes. They are used for classification and explanation. The evaluation shows that the approach finds diverse prototypes. A study involving end users to contrast this approach with other prototype-based methods could potentially lead to new insights regarding the interpretability of the extracted prototypes. Das *et al.* [103] *extract prototypes from the latent representation of the input data* in a deep neural network. The prototypes are weighted, and a surrogate model is built from the prototypes. The approach also applies to data other than time series, yet, a case study for time series data is presented. In this case study, entire time series (stemming from longer time series segmented prior to training the model) are used as prototypes. The authors describe their approach as model-agnostic, since it can be used for different deep learning network architectures. However, following the proposed taxonomy, we categorize it as model-specific since it is constrained to neural networks.

Tang *et al.* propose the *Dual Prototypical Shapelet Networks* [106]. While the paper focuses on few-shot learning, it contributes to interpretability by combining explanation methods offering both local and global explanations. A new representation is learned in the first step, incorporating shapelets and SFA features. The learned representation is then classified with a nearest neighbor classifier. Interpretability is achieved by representative time series examples as well as representative and discriminative shapelets.

While the main motivation of Zhang *et al.*'s approach TapNet [107] is to cope with the lack of labeled data, a contribution to interpretability is also made. For multivariate time series, embeddings are learned by creating subgroups of the univariate time series, followed by 1D-CNNs applied to these groups. In addition, an LSTM is trained on the original multivariate time series. From the learned embeddings, prototypes are extracted and then used for classification. The prototypes are shown in a t-SNE [145] projection of the embedded space.

Summary and analysis. As for improper subsequence prototypes, instance-based prototypes hold valuable information and represent an interesting direction for explainable time series classifiers. However, the black-boxes, usually neural networks, used to build the prototypes are still not explained. Another drawback is that the majority of the works assume that the prototypes are interpretable, focusing solely on evaluating known prototype metrics. For example, the approach in [107] is promising, but the prototypes are visualized with a projection method that is not intrinsically understandable. In some of the studies, the prototypes themselves are analyzed.

Counterfactual-based Explanations. Counterfactual time series show the minimal changes in the input data that lead to a different decision outcome. Formally,

Definition 10 (Counterfactual)**.** Given a classifier f that outputs the decision $y = f(x)$ for an instance x, a *counterfactual* consists of an instance x' such that the decision for f on x' is different from y, i.e., $f(x') \neq y$, and such that the difference between x and x' is *minimal*, and that x' is *plausible*.

Minimality and plausibility depend on the domain where counterfactuals are necessary. However, for time series, minimality typically refers to a notion of distance between time series. On the other hand, plausibility refers to notions involving the usage of outlier detection metrics or measuring the presence of anomalies in the time series.

In [113], Delaney *et al.* propose to extract potential counterfactual time series, named *native guides*, from initial training data. As a first step, these are real time series belonging to D. These are then adapted to generate novel counterfactuals, following four identified key properties for good counterfactuals: proximity, sparsity, plausibility, and diversity. While these properties are generally assumed to yield interpretable examples (prototypes or counterfactuals), evaluating this promising approach involving end users could be promising for future work. The approach is quantitatively evaluated for these properties and compared to two benchmark methods.

In [114], Karlsson *et al.* define the problem of locally and globally explainable time series tweaking. They propose the two approaches τ_{RT} and τ_{IRT} that try to find the minimum amount of changes to a time series that forces a classifier into changing its classification output. The authors focus on implementing the approach for two classifiers, namely k-Nearest Neighbors and Random Shapelet Forest [146]. Evaluating the interpretability of the random forest could be future work.

In [115], Ates *et al.* propose *CoMTE*, an explainability method that provides explanations for multivariate time series classification in terms of counterfactuals. Using a heuristic search algorithm, CoMTE finds a distractor time series from the training set and computes the minimal number of substitutions in order to change the class of the original time series to that of the distractor.

Labeien *et al.* [116] transferred the concept of Contrastive Explanation Method (CEM) [147] to time series classification. Using an LSTM in combination with a fully connected network, they find the minimal perturbations for the model to change its classification decision. These perturbed instances are called pertinent negatives, being similar to counterfactuals. Note that the model is reported to be model-agnostic, while we view it as model-specific since it relies on a combination of LSTM and an autoencoder.

Summary and analysis. Counterfactual-based explanations are grounded in human explanation theory and are intuitively understandable [22]. However, generating counterfactuals for time series is not as trivial as in traditional approaches like [148] that use gradient-based optimization. Those approaches bear the risk of generating time series counterfactuals that are not consistent with the original data [113], i.e., adversarial examples. Thus, a counterfactual-based explanation can be hard to generate but can potentially be a human-acceptable explanation for time series classifiers.

Summary of instance-based methods. There is less research on instance-based XAI methods compared to time points- and subsequences-based methods. The introduced methods comprise prototypes, counterfactuals, as well as feature-based methods. The vast majority of the methods were originally designed for time series data. We would like to point out that there are more interesting feature-based methods that we did not include due to domain-specific feature extraction.

D. OTHER TYPES OF EXPLANATIONS

We place in this section a miscellanea of methods that, in our opinion, do not return explanations that can be easily assigned to one of the previous categories.

In [117], Okajima and Sadamasa propose *Rule-Constrained Network (RCN)*, i.e., a neural network trained to make decisions by selecting decision rules. Given an instance, RCNs select a decision rule from a given set so that the observation satisfies the antecedent of the rule and the consequent gives a high probability to the correct class. An RCN is a data-agnostic model, but in [117] is successfully applied on time series after learning the rules on a TSC dataset.

While Huang *et al.* [118] focus on anomaly detection under specific assumptions, we include it due to its interesting and completely different approach. They address the detection of faults in aircraft systems with the underlying assumption that the correlation between the signals in the multivariate time series is stationary in normal operation mode. In fact, a change in the correlation is an indication of a fault. They use *Granger causal graphs to learn and represent the causalities between the signals* and classify the data based on the differences between the causalities. Interpretability is then achieved by visualizing the differences in the correlations.

In [119], Mohammadinejad *et al.* propose to learn *Temporal logic formulas* to classify time series. Regarding interpretability, the approach favors shorter formulas, which are viewed as easier to understand. Relating this approach to the common terminology of XAI, we view it as a global explanation – since the temporal logic formulas are determined from the entire training set. In addition, we view the approach as model-specific, since the learned formulas are used as a classification model. In contrast to the vast majority of the reviewed papers, this approach does not present the explanations visually, but rather as temporal logic. The authors note that general temporal logic can be transformed into plain text explaining decisions. This is, however, not shown for the approach in the paper and might be a promising future research.

Summary of other methods. This subsection comprised promising approaches for future research directions, such as temporal logic, rules, or Granger causality analysis. However,

these approaches are quite different from the typical work on XAI and might not be within the scope of XAI researchers. Although these works may not be directly accessible to XAI researchers due to the use of different methods and terminology, we would like to highlight their promising ideas.

VI. APPLICATIONS AND EVALUATION

Explainable AI strives in many different research fields and proposes solutions to overcome uncertainties with explainability in high-stakes applications. Various *applications* present use cases of how to deploy machine learning models and apply XAI to increase the understanding and knowledge of the underlying model and gain the users' trust. Further, *evaluation* is a key performance indicator of how reasonable these explanations are on the model and the data for the tackled problem.

A. APPLICATIONS

Due to the tremendous amounts of data generated by an even larger number of deployed sensors, time series classification can be applied in a nearly endless amount of tasks, such as *anomaly detection* for cardiovascular [149] and brain diseases [150], human *activity recognition* [151], *pattern extraction* [152], real-time *crash prediction* [153] and so on. In the past, fewer machine learning models were applied in critical tasks, given their insufficient performance and/or the need for understandable decision-making. Recently, due to the huge success of deep learning models in fields such as computer vision, more and more machine learning algorithms are tested and applied to automate and solve problems. XAI helps to further increase such models' reach into areas where there is a need for understandable decisions. Without claiming to be exhaustive, in the following paragraphs we present examples of some relevant applications.

Technical Systems. Technical systems typically incorporate machine learning models to automate or improve existing tasks previously solved by humans or handcrafted algorithms, for example, in order to identify critical events in production [8] or to diagnose bearing faults [154]. Explainability for TSC has been used in various papers on power consumption. In [140] *shapelets* are used to classify different events in power consumption. The authors report that through these subsequences, they are able to investigate the data, understand which shapelet leads to an event, and hence improve the trust in the system. The investigation is then used to understand which appliance caused an occurred event. Other works like [65] uses a CNN to classify power consumption of different households and use *attribution* techniques to show relevant features and time points for the classification. Their visualization shows a heatmap of the important values. Furthermore, predictive maintenance is a field where time series are commonly encountered. In [155], a 1D-CNN is used on univariate vibration signals to classify faults in linear motion guides. The explanation is achieved using *Grad-CAM* in the frequency domain, arguing that different errors manifest themselves through different frequency patterns. An LSTM is used in [156] to classify rolling-element bearings. The explanation method used is *LRP* to show a heatmap in the time domain. Because time series data is present in a wide range of technical systems, such as manufacturing, automotive, or the internet of things, we consider this application field highly relevant and demanding for interpretable time series classification.

Medical Domain. In the medical domain, it is even more critical to support the trust of doctors and specialists in automatic decision systems in their daily work. In [157], a framework is presented to apply *post-hoc attribution* methods such as *LIME* in medical time series to explain the decisions of models. They apply a DNN on ECG data to classify patients' conditions and use LIME as an explanation extractor to show users how the classification was achieved. Many others also apply attribution methods on ECG [69] or robotics data on surgical tasks [158] to extract explanations from models. In [159], a CNN with an *attention* mechanism is utilized for epileptic seizure detection in multivariate EEG time series. The approach extracts the importance of each EEG channel (one signal of the multivariate time series). Furthermore, the authors show how the attention mechanism allows locating the important brain regions based on the location of the electrodes. Furthermore, for seizure detection, [160] applied *SHAP* on multivariate EEG data to identify important EEG channels. The need for interpretable models is rather obvious in the medical domain. Often, applications of XAI approaches in this domain are indeed *steps towards deployment in a real-world scenario.*

B. EVALUATION

Evaluations are crucial to validate working approaches for practical computer science problems. However, due to different evaluation methodologies, there is not a straightforward and standard solution to apply to every challenge. In most cases, evaluations can be split into two different types: *quantitative* and *qualitative*. Quantitative evaluation metrics focus on evaluating the performance of a model, comparing it to other approaches, for example, measuring their accuracy on a benchmark dataset. On the contrary, qualitative evaluations do not strictly focus on just a measure and use subjective human decisions to measure performances. Humans are often involved in interpreting results giving opinions on which proposed solution works best for a specified task, for example, by looking at the realism of the images generated by a model.

XAI inherits these two approaches to evaluate explanations from the field of computer science, as these are highly subjective towards the target user group. For example, engineers working with time series and forecasting models need a different explanation than maintenance workers repairing engines and maintaining large production machines. Thus, *extracting the proper explanation for each group involves different evaluations*. At first, the XAI technique needs to be evaluated, and then a suitable medium for the explanation

needs to be found and presented to users [161].

Typically, a first initial analysis of the explanations is undertaken in a visual assessment of individual samples with experts. Afterward, in some cases, a broader range of experts and users is included to evaluate the explanations. To generally demonstrate that the XAI approaches are working, quantitative fidelity analysis is used to assess the trustworthiness of the model's explanation. Then in some cases, enhancements of the fidelity or other quantitative and automatic evaluation measures are used to get further insights into the approach and the classifier. We believe that the evaluation of XAI, in general, is challenging and currently not solved at a satisfactory level. The lack of widely accepted metrics or test procedures for the young research field of XAI for time series classification is even more evident. For general XAI, the position paper [42] made a strong case for rigorous evaluation methods. These thoughts should also be taken into account in the time series domain. Currently, there is no accepted metric to quantify the interpretability of the reviewed methods, i.e., user studies will be necessary for the time being. In addition, it is desirable to have a set of quantified metrics since they are less subjective and easier to compare. We believe a combination of quantitative results and user studies will be necessary to convince potential users of the proposed XAI methods.

In the following, we shortly introduce the mentioned concepts:

Visual Evaluation. Beginning with qualitative evaluation, a first initial judgment is often done by inspecting the explanation's visualization. In such cases, visualizations facilitate presenting an explanation to users [120]. In Table 4 we refer to this type of explanation as author-selected examples. Such first demonstrations give insights into the model as well as into the data used [61], [83], [101], [162]. However, these *visual approaches are highly qualitative evaluations* given that in most cases, only small-scale studies with a limited amount of users are undertaken. When the evaluation involves only domain experts, the scale is even smaller, and the measurement of their understanding of a model's behavior and explanations is limited [62], [105]. However, these initial evaluations often lead to either feedback for further research opportunities or present valuable empirical data to support the claims of applicable state-of-the-art approaches [62], [122]. The main drawback is that such visual evaluations, being rather subjective, can possibly lead to faulty conclusions. Hence, they should be verified together with a quantitative method [163].

Faithfulness Analysis. The most prominent and widely used quantitative evaluation for XAI in TSC is the *faithfulness analysis*. Faithfulness describes how accurate an explanation fits a model's behavior towards the prediction score. Throughout literature, faithfulness is referred to under various terms, for instance perturbation analysis [83], ablation study, trustworthiness, or fidelity [61]. In many cases, such a fidelity analysis is achieved by explaining a sample and changing (perturbing) the relevant parts from the explanation of the sample to non-informative values [7]. Afterward, the change in prediction is observed either over just one sample [111] or other a whole set of samples [7]. In cases over more than one sample, a quality metric presents changes in the prediction. Often, a significant change in the quality metric score is assumed to highlight good working explanations [7].

Quality metrics that are applied in such cases often use the underlying training metrics such as accuracy (ACC) or area under the receiver operating characteristic curve (AUROC). Examples for such evaluations using accuracy are [7], [62], [83], [111], [162], AUROC [67], [68], AUPRC (area under the precision-recall curve) [67], [164].

After selecting a quality metric to observe, the perturbation analysis can either remove or retain relevant values in samples for the observation [165]. However, removing and retaining information still needs a non-information holding value as a baseline to perturb the data. In time series, such values are not easy to identify for many datasets [166], for example, values like 0 could have a specific meaning altering the class membership. Thus, [166] proposes seven different alternatives on how to perturb the data in relevant parts to enable non-information holding values. Besides the focus on individual time points, they also propose perturbation strategies to evaluate the time constraints. Both, [162] and [7], suggest that there is not one best explanation method to use for all models, but rather a heavy dependence on the applied model's approach.

Other Approaches. In [165], *fidelity analysis* is enhanced with a referee to compare the prediction of the initial model with referee classifiers, gradually removing or retaining the most important features highlighted by the explanation. The informativeness of an explanation is assessed by looking at the degradation in performance of the referee classifiers. They argue that *such a referee helps to better compare various XAI techniques against each other*. In [164] a benchmark for XAI techniques is proposed using synthetic data with properties to evaluate the explanations of such methods. In other fields, such as computer vision, there are further techniques to evaluate explanations, e.g., [167] or ROAR [168]. However, most of them are not exhaustively applied to XAI for TSC yet.

VII. DISCUSSION AND RESEARCH DIRECTIONS

During the analysis of the selected papers, we identified various trends and challenges in XAI research for TSC. In the following, we shortly discuss our perspective of the review on these challenges and highlight future research opportunities to close the gaps we identified.

Discussion. We found some interesting relationships between different categories of the proposed taxonomy. From Table 3, it becomes obvious that the granularity of the explanation is related to the locality or globality of the XAI approaches. For example, point-based explanations are mostly local, clarifying the model explanation for individual instances. In contrast, subsequence and instance-based explanations are more often global, shedding light on the whole model's behavior. Next,

TABLE 4. Overview of evaluations conducted in the reviewed papers. The lack of user studies to evaluate interpretability becomes obvious. (Frequently used evaluation methods are shown with the abbreviations: ACC = classification accuracy, EX = author-selected examples and visual presentation, PERTURB = pertubation analysis, RT = runtime, US = user study, QUANTCOMP = quantitative comparison to other methods, POINT = pointing game in localization)

XAI method	Ref	Evaluation in the paper	Comment
INTGRAD	[53]	EX	not shown for time series
FCN	[54]	ACC, EX	
LIME	[36]	EX, US	not shown for time series
LRP	[55]	EX, PERTURB	not shown for time series
ExcitationBP	[56]	EX, POINT	not shown for time series
Occlusion	[57]	ACC, EX, RT	not shown for time series
SHAP	[37]	RT, US	not shown for time series
SmoothGrad	[58]	EX	not shown for time series
DeepLIFT	[39]	ACC, PERTURB	not shown for time series
Salience-CAM	[59]	ACC, EX, PERTURB	
Grad-CAM	[60]	ACC, EX, US	not shown for time series
TSViz	[61]	EX, discussion w.r.t. defined properties	
TSXplain	[62]	EX, PERTURB, US	textual explanations with user study
TSInsight	[63]	ACC, QUANTCOMP	
SoundLIME	[64]	ACC, EX, agreement of saliency map	
MTEX-CNN	[65]	ACC, EX	
PERT	[66]	EX, QUANTCOMP	
FIT	[67]	PERTURB	
WinIT	[68]	PERTURB	
CEFEs	[69]	EX, compared to domain-specific features	domain-knowledge incorporated in eval
XTF-CNN	[70]	EX, compared to domain-specific features	domain-knowledge incorporated in eval
LEFTIST	[71]	PERTURB to black- and white-box models, US	extensive user study (194 participants)
ALSTM-FCN	[12]	ACC	interpr. not evaluated
GCRNN	[72]	ACC, EX	
-	[73]	ACC, EX, RT	
ETSCM	[74]	ACC, earlyness, EX, expert US	evaluated for ECG data by medical doctors
DACNN	[75]	ACC, EX	
LAXCAT	[76]	ACC, RT, QUANTCOMP	
DeepVix	[77]	EX	
VixLSTM	[78]	EX	
-	[79]	EX, ROAR	non-standard quantitative evaluation
-	[10]	ACC	
SAX-VSM	[80]	ACC, EX, RT	early work in the field; eval of interpr. could be future work
-	[81]	ACC, EX visualized as network plot	interpr. of plot could be future work
MR-SEQL	[82]	ACC, EX, RT	focus on ACC, eval of interpr. could be future work
CPHAP	[83]	PERTURB, EX also from other approaches	
Shapelets	[131]	ACC	focus on efficient shapelet discovery
ShapeletTransform	[85]	ACC	focus on efficient shapelet discovery
MSD	[86]	ACC, RT	focus on efficient shapelet discovery
LS	[87]	RT, accuracy of shapelet discovery	focus on efficient shapelet discovery
LTS	[88]	ACC, RT	focus on efficient shapelet discovery
LCTS	[89]	ACC, RT	focus on efficient shapelet discovery
-	[90]	ACC, EX, RT	
-	[91]	ACC	
LRS	[92]	RT, accuracy of shapelet discovery	
ADSNs	[93]	ACC, EX	
XCNN	[94]	ACC, EX	
GENDIS	[95]	ACC, RT, accuracy of shapelet discovery	focus on ACC
GMSM	[96]	ACC	
MAPIC	[97]	ACC, RT	
DASH	[98]	ACC, RT, faithfulness, stability	
TORRENT	[99]	ACC, EX, RT	
LASTS	[100]	EX, faithfulness, stability, QUANTCOMP	
PatchX	[101]	ACC, EX, RT	uses variable-sized patches and evaluates an own patch-based metric
P2ExNet	[102]	ACC, PERTURB	
ProtoFac	[103]	EX, fidelity	contains user study, but not for time series data
mWDN	[104]	ACC, EX	
ProSeNet	[105]	US	contains user study, but not for time series data
-	[52]	ACC, prototype diversity	evaluated in four brief case studies
DPSN	[106]	ACC, EX	
Tapnet	[107]	ACC, projection vis of dataset with prototypes	
MODL-TSC	[108]	EX	early work in the field; eval of interpr. could be future work
MTDT	[109]	ACC, EX	
FC/TAA/LTAA	[110]	ACC, RT	focus on computational efficiency
Conceptual	[111]	ACC, EX, manual distortion of data (occlusion-like)	
-	[112]	EX	
Native Guide	[113]	QUANTCOMP	eval of four metrics for counterfactuals
τ_{RT}/τ_{IRT}	[114]	ACC, RT	interpr. of created random forest could be future work
CoMTE	[115]	w.r.t. known feature importance	domain-knowledge incorporated in eval
CEM	[116]	ACC, manual distortion of data	
RCN	[117]	ACC, EX	
TCCL	[118]	EX	
-	[119]	EX	promising, uncommon method; eval of interpr. could be future work

we identified connections between the explanation method and ante/post-hoc approaches. In particular, attribution methods are primarily post-hoc, working on fixed trained models without the need to make assumptions about the explained classifier. On the other hand, attentions and shapelets are built into the model as a constraint and an ante-hoc mechanism, with interpretability directly embedded in the approach. Instance-based methods are more heterogeneous, with a similar amount of ante/post-hoc approaches. From the reviewed papers, model-agnostic approaches are rarer compared to model-specific approaches and are, for the most part, local. Finally, an interesting observation is that time-point-based methods have been proposed for univariate and multivariate time series about equally often, while for subsequence- and instance-based methods, the focus is on univariate time series, with just a few approaches for multivariate time series. While the use of subsequences is more challenging, there appears to be a research gap due to the high practical relevance of multivariate time series.

It is clear from the publication dates that subsequences-based methods, such as SAX or shapelets, have a rich history with many solid methods. However, in that field, computational efficiency is frequently given more attention.

Since we believe that the evaluation of XAI methods is challenging and has not yet reached a satisfactory state [7], [42], we analyzed how the XAI methods proposed in the reviewed papers were evaluated (see Table 4): evaluation is quite frequently done using author-selected examples, showing that the approach is plausible. As a quantitative method, classification accuracy is frequently used, showing, for example, that explainable methods are competitive with their non-explainable counterparts. Accuracy, however, can not assess the methods' interpretability. To a much lesser degree, other quantitative explanation measures are used. User studies evaluating the interpretability were presented in a small minority of the papers. Interestingly, none of the reviewed subsequences-based approaches was validated with a user study. We view the lack of user studies as an important observation that, from our point of view, points to a deficiency in a research field that aims to make machine learning interpretable for human beings. One strong underlying motivation of XAI is to enable users to trust machine learning models that, without explanation, have a black-box nature. Not involving users in evaluating explanations bears the risk that proposed explanations are not accepted for use in practical applications for the same reasons as not accepting machine learning in the first place. Hence, we view the involvement of users as necessary for future work until the interpretability of a new method is shown. Further work, such as refining or making a method more efficient, may not then require user studies. Also, if quantitative methods to measure interpretability are found and widely accepted in the field, the incorporation of users might not be required for each study.

While for machine learning on feature vectors, decision trees, rule bases, or – for a limited number of features – linear models are generally assumed to be intrinsically interpretable[1], we do not see analogous models for time series. While decision trees have been used for time series, they rely on some feature extraction step or the use of shapelets. The entire process can not simply be considered interpretable. Distance-based methods might, to some degree, be considered interpretable; however, they suffer the same problem for more complex distance measures.

Regarding code implementation, we found that the vast majority of the published work uses Python as a programming language. The most famous and used libraries for time series are those with good documentation and implementation, for example `Shap` [37], `Captum` [169], `GradCAM` [170], `Lime` [36] for time point-based explanations, `sktime` [171], `tslearn` [172] and `pyts` [173] for subsequences-based explanations. At the same time, some promising approaches are not as widespread, given their lack of a simple and easy-to-use codebase. We believe that a library specifically designed for explaining time series classifiers is still missing.

As a critique of our taxonomy, we observe that a small subset of papers, not aligned with typical XAI works, could not be classified under it; for this reason, the "others" subcategory was added. In addition, some approaches could be considered hybrid (e.g., time-points and subsequences-based), and this was indicated in the text. Furthermore, we acknowledge that our selection of reviewed papers might have missed interesting work. For example, we did not include preprints and Ph.D. theses, which might also hold compelling ideas.

Research directions. Based on the reviewed papers and our work in the field, we identified a number of research directions that we believe can contribute to inspire research in the field.

a: Higher-order explanations are desirable

Throughout reviewing our selected papers, we often observed visualizations showing the time series with a heatmap on top of the line plot or behind it, as described in [120] for attributions or attentions. For subsequences and some applications, we mainly observed a highlighting of the relevant part of the time series corresponding to the subsequence. However, we argue that such visualizations are not sufficient in those cases where the pure signal cannot be directly interpreted. In contrast to computer vision, highlighted parts of a time series are not directly interpretable for all problem settings. We see an opportunity for higher-order representations besides line plots of the explanations to enable a more straightforward explanation. Based on such other representations, we further highlight the importance of explanations that are not purely visualizations. Verbalization, i.e., textual descriptions of the explanation, can also explain the decision and behavior of the black-box model in the specific terminology of the problem domain [174] enabling experts to understand the model better. An example of a textual explanation is TSXplain [62] where the explanations were found to be valuable in a user study. A

[1]Note that also decision trees and rule bases can become too complex to be easily interpretable.

combination of verbalization and time series visualizations can help explain models in user terms without the need for in-depth knowledge in the time series domain regarding properties like, e.g., periodicity or structure of the data, understanding of time series representations, and algorithms. While machine learning engineers often use XAI methods, explanations offered to end users should be sufficient to understand the application domain itself, e.g., manufacturing, automotive, or medical domain. In [49], this issue is referred to as quasi-explanations which are explanations containing items that are foreign to the domain.

b: Model-agnostic approaches are particularly useful for TSC

For TSC, a variety of different model types is used. In contrast to, e.g., computer vision applications, deep learning methods do currently not clearly dominate the field. In the search for the best method for a given problem setting, deep learning [29], ensembles [28], distance-based methods [32], shapelets and further methods are used. In order to compare the interpretability of these entirely different model architectures, model-agnostic methods are required. Model-specific methods may then be used at a later stage of the model selection process.

c: Domain-specific explanations for specific applications

In general, building models and explanations that work in a wide range of fields is desirable. However, we believe some cases require domain-specific explanations when explaining the models to end users. Indeed, the effectiveness of an explanation depends on the user's perception and response rather than on the model. In particular, the end user may not be able to understand all the information even if a model is made entirely transparent. In other words, given that explanations depend on the requirements of the target users, there is no one-size-fits-all solution in the expanding body of XAI techniques: what makes an explanation effective depends on the user's goals, background, and current level of knowledge [175]. When addressing these issues, differentiating user groups is a good place to start. For instance, machine learning experts might want to enhance or debug deep learning models, business owners might want to assess compliance with regulations, and laypeople might want to gauge how AI decisions affect their daily lives. Furthermore, XAI approaches can be improved by considering the application domain while incorporating domain-specific knowledge, yielding useful explanations instead of quasi-explanations [49]. For instance, in the healthcare industry, doctors would greatly benefit from XAI techniques that could clarify the AI diagnosis and enable the injection of the expert's knowledge to enhance the quality of the explanations [176], [177]. Domain knowledge is also essential in AI-based cybersecurity systems, where explanations must satisfy the needs of many stakeholders [178]. Besides domain-specific XAI methods, a research direction might be to develop *general* XAI methods that incorporate *specific* domain knowledge. Finally, tighter integration of humans into the explanation generation can lead to a better evaluation of the methods.

d: Easy-to-use explainers are desirable

Some XAI methods to explain black-box models might be viewed as black-boxes themselves, as pointed out in previous works [27]. Explaining complex ML models calls for sophisticated and often complicated XAI methods. For example, some XAI methods come with assumptions like local linearity or feature independence (see [123], [124] for a discussion) and many hyperparameters, e.g., to control approximations of computationally expensive explanations. For these reasons, there is a risk that the XAI methods will produce invalid explanations.

e: More rigorous evaluation of explanations is needed

We view the evaluation of explanations as a crucial challenge in XAI research in general, and specifically in XAI research for time series. Evaluation of explainability in other domains such as computer vision is much more advanced than in the time series domain. In particular, adapting some of these evaluation techniques to previously presented XAI methods can lead to first insights into developing more specialized ones. In general, further methods for faithfulness analysis need to be implemented and generally be used to establish a rigorous quantitative and automatic evaluation. Finally, focusing on verbalization and not only visualization can help improve such evaluations even more, as possibly faulty explanations can be identified more easily.

f: Evaluation should also address human interpretability

We observed that research in the field often addresses the part of explainability, i.e. explaining a model by means of e.g. a visualisation, shapelets or prototypes. These artefacts are then shown with author-selected examples or in some cases evaluated quantitatively. Whether the artefacts are indeed interpretable for the target audience is an evaluation that is not regularly conducted. The artefacts are, by definition, assumed to be interpretable. However, as, e.g., shapelets can be different to the initial dataset time series samples (see examples in [142]), users do not necessarily understand such explanations. The same is true in the case of highlighting parts of the input data that do not correspond with the understanding of domain experts (as shown in a study with medical doctors in [74]). Furthermore, extracted prototypes are typically evaluated with respect to acknowledged properties for prototypes (e.g., sparsity, diversity). However, whether or not these prototypes are useful to end users is not always evaluated. This fact emphasizes the need to involve end users in evaluating a method's interpretability, as also stressed in [49]. We believe that quantitative metrics should be used to support user studies or, if progress is made in studying the human interpretability of XAI explanations, perhaps even to replace them in the future.

g: Unified implementation of XAI for comparative evaluation

While many of the reviewed XAI methods provide source code (see Table 3), there is no unified library that allows for easy comparative evaluation of XAI methods. Looking back

at the history of research on time series, we believe that the initiative by Keogh and other researchers to provide a unified data archive [179] for evaluating time series algorithms has inspired research in the field and made more rigorous research possible. For time series classification, several programming libraries exist (e.g., `sktime`, `tsai`). An analogous library of XAI methods for time series classification is desirable, ideally in connection with datasets. Recently, promising work towards an XAI benchmark, not specifically aimed at time series, was published as a preprint [180].

h: Benchmark data sets for evaluation are desirable

Quantitatively evaluating an explanation is challenging. Using perturbation methods [7] requires setting thresholds and to alter the original data using some pre-defined values or some obtained background noise. Hence, the evaluation process has a number of parameters itself, which might lead to different results when being used by different researchers. An idea could be to have gold-standard datasets for the evaluation of explanations: time series that are annotated, i.e., subsequences, data points or higher order features that are known to be discriminative are annotated. Unfortunately, such datasets are not available yet. Indeed, typically studies proposing novel XAI approaches for time series experiment on common benchmarking datasets used for TSC[2], not containing annotations regarding interpretability, rather one label per time series. For example, in computer vision there are data sets that have annotations for parts within the image, e.g., the CUBS data set [181] or CLEVR-XAI [182]. To be as complete as possible, it might be an option to synthetically create such datasets, inducing known class-discriminative artefacts into the data. A drawback of such a solution would be a potential overfitting of explanations methods towards gold-standard data sets. Hence, those datasets could be an additional option to check the plausibility of an explanation but not a replacement for the more generic evaluation methods currently being developed e.g., [7], [165], [166]. Furthermore, they will not replace user studies to evaluate the interpretability for the target audience.

VIII. CONCLUSION

In this review, we presented the first extensive overview of the current body of literature regarding XAI for time series classification. We proposed a taxonomy based on the granularity of the explanation, categorizing the reviewed methods into three groups of approaches: time points-, subsequences-, and instance-based. We further highlighted the main approaches to evaluate explanations and the practical challenges of developing quantitative and qualitative metrics towards human and automatic techniques. To inspire further research in the field, we identified various research directions. Specifically, we believe there are research gaps in the fields of higher-order explanations, model-agnostic approaches, domain-specific explanations, easy-to-use explanations, more advanced evaluation of explanations, evaluation of interpretability as well as a unified framework with XAI methods for time series classification and benchmark data sets for their evaluation.

Explainability is a fast-growing subject in the literature, and it is clear that the interest on the topic is rising. XAI approaches for time series data are helpful in building trust towards the decisions of machine learning algorithms, to better support experts and their accountability and responsibility in the decision-making, bringing insights in many critical domains.

[2]Examples are the UCR archive [179] and the Time Series Classification Website.

ACKNOWLEDGMENT

We would like to thank Philip Ritzer and Felix Gerschner for their support. Further, this work has been partially supported by the European Community H 2020 programme under the funding schemes: G.A. 871042 *SoBigData++*, G.A. 952026 *HumanE AI Net*, ERC-2018-ADG G.A. 834756 *XAI: Science and technology for the eXplanation of AI decision making* (`xai`), G.A. 952215 *TAILOR* (`tailor`), CHIST-ERA grant *CHIST-ERA-19-XAI-010*, by MUR (N. not yet available), FWF (N. I 5205), EPSRC (N. EP/V055712/1), NCN (N. 2020/02/Y/ST6/00064), ETAg (N. SLTAT21096), BNSF (N. KP-06-AOO2/5) and by the Federal Ministry of Education and Research (BMBF) in the VIKING (13N16242) project, EXPLOR-20AT by Stiftung Kessler + CO für Bildung und Kultur. Open Access fee funded by Aalen University of Applied Sciences.

REFERENCES

[1] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5):206–215, 2019.
[2] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5):1–42, 2018.
[3] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access, 6:52138–52160, 2018.
[4] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. explAIner: A visual analytics framework for interactive and explainable machine learning. IEEE Transactions on Visualization and Computer Graphics, 26(1):1064–1074, 2019.
[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009.
[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. University of Toronto, 2012.
[7] Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, Daniela Oelke, and Daniel A Keim. Towards a rigorous evaluation of XAI methods on time series. In 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pages 4197–4201. IEEE, 2019.
[8] R Keith Mobley. An introduction to predictive maintenance. Elsevier, 2002.
[9] Andreas Theissler. Detecting known and unknown faults in automotive systems using ensemble-based anomaly detection. Knowledge-Based Systems, 123:163 – 173, 2017.
[10] André Maletzke, Huei Lee, Gustavo Batista, Solange Rezende, Renato Machado, Richardson Voltolini, Joylan Maciel, and Fabiano Silva. Time Series Classification using Motifs and Characteristics Extraction: A Case Study on ECG Databases. In Proceedings of the Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support, 2013.

[11] Kyoung-jae Kim. Financial time series forecasting using support vector machines. Neurocomputing, 55(1-2):307–319, 2003.

[12] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. LSTM Fully Convolutional Networks for Time Series Classification. IEEE Access, 6:1662–1669, 2018.

[13] Andreas Theissler, Judith Pérez-Velázquez, Marcel Kettelgerdes, and Gordon Elger. Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. Reliability Engineering & System Safety, 215:107864, 2021.

[14] Timo Markert, Sebastian Matich, Elias Hoerner, Andreas Theissler, and Martin Atzmueller. Fingertip 6-Axis Force/Torque Sensing for Texture Recognition in Robotic Manipulation. In 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), pages 1–8. IEEE, 2021.

[15] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. Explainable AI: the new 42? In International cross-domain conference for machine learning and knowledge extraction, pages 295–303. Springer, 2018.

[16] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44):22071–22080, 2019.

[17] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8):832, 2019.

[18] Christoph Molnar. Interpretable Machine Learning. Lulu. com, 2020.

[19] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. Explainable AI: interpreting, explaining and visualizing deep learning, volume 11700. Springer Nature, 2019.

[20] David Martens, Bart Baesens, Tony Van Gestel, and Jan Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. Eur. J. Oper. Res., 183(3):1466–1476, 2007.

[21] Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. Explainable artificial intelligence: A survey. In MIPRO, pages 210–215. IEEE, 2018.

[22] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell., 267:1–38, 2019.

[23] Ruth M. J. Byrne. Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning. In Sarit Kraus, editor, IJCAI, pages 6276–6282. ijcai.org, 2019.

[24] Yongfeng Zhang and Xu Chen. Explainable recommendation: A survey and new perspectives. Found. Trends Inf. Retr., 14(1):1–101, 2020.

[25] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In AAMAS, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[26] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): towards medical XAI. CoRR, abs/1907.07374, 2019.

[27] Simon Vollert, Martin Atzmueller, and Andreas Theissler. Interpretable Machine Learning: A brief survey from the predictive maintenance perspective. In IEEE International Conference on Emerging Technologies and Factory Automation (ETFA 2021). IEEE, 2021.

[28] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery, 31, 05 2017.

[29] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. Data Min. Knowl. Discov., 33(4):917–963, 2019.

[30] Alejandro Ruiz, Michael Flynn, James Large, Matthew Middlehurst, and Anthony Bagnall. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery, 35:1–49, 03 2021.

[31] Lamyaa Sadouk. CNN approaches for time series classification. In Time Series Analysis-Data, Methods, and Applications, pages 1–23. IntechOpen, 2019.

[32] Amaia Abanda, Usue Mori, and José Antonio Lozano. A review on distance based time series classification. Data Min. Knowl. Discov., 33(2):378–412, 2019.

[33] Paolo Tormene, Toni Giorgino, Silvana Quaglini, and Mario Stefanelli. Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. Artificial intelligence in medicine, 45(1):11–34, 2009.

[34] Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey. CoRR, abs/2104.00950, 2021.

[35] Ilija Simic, Vedran Sabol, and Eduardo E. Veas. XAI methods for neural time series classification: A brief review. CoRR, abs/2108.08009, 2021.

[36] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144, 2016.

[37] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems, pages 4768–4777, 2017.

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pages 618–626, 2017.

[39] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 3145–3153. JMLR.org, 2017.

[40] Aaron Bostrom and Anthony Bagnall. A shapelet transform for multivariate time series classification. 12 2017.

[41] Alex Alves Freitas. Comprehensible classification models: a position paper. SIGKDD Explor., 15(1):1–10, 2013.

[42] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. arXiv, 2017.

[43] Luca Longo, Randy Goebel, Freddy Lécué, Peter Kieseberg, and Andreas Holzinger. Explainable artificial intelligence: Concepts, applications, research challenges and visions. In CD-MAKE, volume 12279 of Lecture Notes in Computer Science, pages 1–16. Springer, 2020.

[44] What Do People Really Want When They Say They Want Explainable AI? We Asked 60 Stakeholders, New York, NY, USA, 2020. ACM.

[45] David Gunning and David Aha. DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine, 40(2):44–58, 2019.

[46] Nicholas Diakopoulos, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, H. V. Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, Christo Wilson, Cong Yu, and Bendert Zevenbergen. Fairness, Accountability, and Transparency in Machine Learning, 2018.

[47] Or Biran and Courtenay V. Cotton. Explanation and Justification in Machine Learning : A Survey. In IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), 2017.

[48] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58:82–115, 2020.

[49] Boris Kovalerchuk, Muhammad Aurangzeb Ahmad, and Ankur Teredesai. Survey of Explainable Machine Learning with Visual and Granular Methods Beyond Quasi-Explanations, pages 217–267. Springer International Publishing, Cham, 2021.

[50] Zachary C. Lipton. The Mythos of Model Interpretability. Queue, 16(3):31–57, 2018.

[51] Hannah Snyder. Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104:333–339, 2019.

[52] Alan H Gee, Diego Garcia-Olano, Joydeep Ghosh, and David Paydarfar. Explaining deep classification of time-series data with learned prototypes. In CEUR workshop proceedings, volume 2429, page 15. NIH Public Access, 2019.

[53] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 3319–3328. JMLR.org, 2017.

[54] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In 2017 International Joint Conference on Neural Networks (IJCNN), pages 1578–1585, 2017.

[55] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7):e0130140, 2015.

[56] Jianming Zhang, Zhe L. Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-Down Neural Attention by Excitation Backprop. In

Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV, volume 9908 of Lecture Notes in Computer Science, pages 543–559. Springer, 2016.

[57] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I, volume 8689 of Lecture Notes in Computer Science, pages 818–833. Springer, 2014.

[58] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. CoRR, 06 2017.

[59] Linjiang Zhou, Chao Ma, Xiaochuan Shi, Dian Zhang, Wei Li, and Libing Wu. Salience-cam: Visual explanations from convolutional neural networks via salience score. In 2021 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE, 2021.

[60] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision, 128(2):336–359, Oct 2019.

[61] Shoaib Ahmed Siddiqui, Dominique Mercier, Mohsin Munir, Andreas Dengel, and Sheraz Ahmed. TSViz: Demystification of Deep Learning Models for Time-Series Analysis. IEEE Access, 7:67027–67040, 2019.

[62] Mohsin Munir, Shoaib Ahmed Siddiqui, Ferdinand Küsters, Dominique Mercier, Andreas Dengel, and Sheraz Ahmed. TSXplain: Demystification of DNN Decisions for Time-Series Using Natural Language and Statistical Features. Lecture Notes in Computer Science, page 426–439, 2019.

[63] Shoaib Ahmed Siddiqui, Dominique Mercier, Andreas Dengel, and Sheraz Ahmed. TSInsight: A Local-Global Attribution Framework for Interpretability in Time Series Data. Sensors, 21(21):7373, 2021.

[64] Saumitra Mishra, Bob L. Sturm, and Simon Dixon. Local Interpretable Model-Agnostic Explanations for Music Content Analysis. In ISMIR, 2017.

[65] Roy Assaf, Ioana Giurgiu, Frank Bagehorn, and Anika Schumann. MTEX-CNN: Multivariate Time Series EXplanations for Predictions with Convolutional Neural Networks. In 2019 IEEE International Conference on Data Mining (ICDM), pages 952–957, 2019.

[66] Prathyush S. Parvatharaju, Ramesh Doddaiah, Thomas Hartvigsen, and Elke A. Rundensteiner. Learning Saliency Maps to Explain Deep Time Series Classifiers, page 1406–1415. Association for Computing Machinery, New York, NY, USA, 2021.

[67] Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K Duvenaud, and Anna Goldenberg. What went wrong and when? Instance-wise feature importance for time-series black-box models. In Advances in Neural Information Processing Systems, volume 33, pages 799–809. Curran Associates, Inc., 2020.

[68] Clayton Rooke, Jonathan Smith, Kin Leung, Maksims Volkovs, and Saba Zuberi. Temporal Dependencies in Feature Importance for Time Series Predictions. In ICML 2021, Time series workshop. Poster session., 2021.

[69] Barbara Mukami Maweu, Sagnik Dakshit, Rittika Shamsuddin, and Balakrishnan Prabhakaran. CEFEs: A CNN Explainable Framework for ECG Signals. Artificial Intelligence in Medicine, 115:102059, 2021.

[70] Xin Bi, Chao Zhang, Yao He, Xiangguo Zhao, Yongjiao Sun, and Yuliang Ma. Explainable time-frequency convolutional neural network for microseismic waveform classification. Information Sciences, 546:883–896, 2021.

[71] Maël Guillemé, Véronique Masson, Laurence Rozé, and Alexandre Termier. Agnostic local explanation for time series classification. In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pages 432–439. IEEE, 2019.

[72] Sangdi Lin and George C. Runger. GCRNN: Group-Constrained Convolutional Recurrent Neural Network. IEEE Transactions on Neural Networks and Learning Systems, 29(10):4709–4718, 2018.

[73] Phongtharin Vinayavekhin, Subhajit Chaudhury, Asim Munawar, Don Joven Agravante, Giovanni De Magistris, Daiki Kimura, and Ryuki Tachibana. Focusing on what is relevant: Time-series learning and understanding using attention. In 24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, August 20-24, 2018, pages 2624–2629. IEEE Computer Society, 2018.

[74] En-Yu Hsu, Chien-Liang Liu, and Vincent Tseng. Multivariate Time Series Early Classification with Interpretability Using Deep Learning and Attention Mechanism, pages 541–553. Springer International Publishing, 03 2019.

[75] Babak Hosseini, Romain Montagné, and Barbara Hammer. Deep-aligned convolutional neural network for skeleton-based action recognition and segmentation (extended version article). Data Science and Engineering / Issue 2/2020, 2020.

[76] Tsung-Yu Hsieh, Suhang Wang, Yiwei Sun, and Vasant Honavar. Explainable Multivariate Time Series Classification: A Deep Neural Network Which Learns to Attend to Important Variables As Well As Time Intervals, page 607–615. Association for Computing Machinery, New York, NY, USA, 2021.

[77] Tommy Dang, Hao Van, Huyen Nguyen, Vung Pham, and Rattikorn Hewett. DeepVix: Explaining Long Short-Term Memory Network With High Dimensional Time Series Data. In Proceedings of the 11th International Conference on Advances in Information Technology, IAIT2020, New York, NY, USA, 2020. Association for Computing Machinery.

[78] Tommy Dang, Huyen N. Nguyen, and Ngan V.T. Nguyen. VixLSTM: Visual Explainable LSTM for Multivariate Time Series. In The 12th International Conference on Advances in Information Technology, IAIT2021, New York, NY, USA, 2021. Association for Computing Machinery.

[79] Leonid Schwenke and Martin Atzmueller. Show me what you're looking for: visualizing abstracted transformer attention for enhancing their local interpretability on time series data. In The International FLAIRS Conference Proceedings, volume 34, 2021.

[80] Pavel Senin and Sergey Malinchik. SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model. In 2013 IEEE 13th International Conference on Data Mining, 12 2013.

[81] Wei Song, Lu Liu, Minghao Liu, Wenxiang Wang, Xiao Wang, and Yu Song. Representation learning with deconvolution for multivariate time series classification and visualization. In International Conference of Pioneering Computer Scientists, Engineers and Educators, pages 310–326. Springer, 2020.

[82] Thach Nguyen, Severin Gsponer, Iulia Ilie, Martin O'reilly, and Georgiana Ifrim. Interpretable time series classification using linear models and multi-resolution multi-domain symbolic representations. Data Min. Knowl. Discov., 33(4):1183–1222, jul 2019.

[83] Sohee Cho, Wonjoon Chang, Ginkyeng Lee, and Jaesik Choi. Interpreting Internal Activation Patterns in Deep Temporal Neural Networks by Finding Prototypes. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '21, page 158–166, New York, NY, USA, 2021. Association for Computing Machinery.

[84] Lexiang Ye and Eamonn Keogh. Time series shapelets: A novel technique that allows accurate, interpretable and fast classification. Data Min. Knowl. Discov., 22:149–182, 01 2011.

[85] Jason Lines, Luke M. Davis, Jon Hills, and Anthony Bagnall. A Shapelet Transform for Time Series Classification. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, page 289–297, New York, NY, USA, 2012. Association for Computing Machinery.

[86] Mohamed Ghalwash and Zoran Obradovic. Early classification of multivariate temporal observations by extraction of interpretable shapelets. BMC Bioinformatics, 13:195, 08 2012.

[87] Abdullah Mueen, Eamonn Keogh, and Neal Young. Logical-Shapelets: An Expressive Primitive for Time Series Classification. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, page 1154–1162, New York, NY, USA, 2011. Association for Computing Machinery.

[88] Josif Grabocka, Nicolas Schilling, Martin Wistuba, and Lars Schmidt-Thieme. Learning Time-Series Shapelets. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, page 392–401, New York, NY, USA, 2014. Association for Computing Machinery.

[89] Yi Yang, Qilin Deng, Furao Shen, Jinxi Zhao, and Chaomin Luo. A shapelet learning method for time series classification. In 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), pages 423–430. IEEE, 2016.

[90] Zicheng Fang, Peng Wang, and Wei Wang. Efficient learning interpretable shapelets for accurate time series classification. In 2018 IEEE 34th International Conference on Data Engineering (ICDE), pages 497–508. IEEE, 2018.

[91] Huiqi Deng, Weifu Chen, Andy J. Ma, Qi Shen, Pong C. Yuen, and Guocan Feng. Robust shapelets learning: Transform-invariant prototypes. In Jian-Huang Lai, Cheng-Lin Liu, Xilin Chen, Jie Zhou, Tieniu Tan, Nanning Zheng, and Hongbin Zha, editors, Pattern Recognition and Computer Vision, pages 491–502, Cham, 2018. Springer International Publishing.

[92] Maël Guilleme, Simon Malinowski, Romain Tavenard, and Xavier Renard. Localized Random Shapelets. In International Workshop on Advanced Analysis and Learning on Temporal Data, pages 85–97, Wurzburg, Germany, 2019.

[93] Qianli Ma, Wanqing Zhuang, Sen Li, Desen Huang, and Garrison Cottrell. Adversarial Dynamic Shapelet Networks. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04):5069–5076, Apr. 2020.

[94] Yichang Wang, Rémi Emonet, Elisa Fromont, Simon Malinowski, and Romain Tavenard. Adversarial Regularization for Explainable-by-Design Time Series Classification. In 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), pages 1079–1087, 2020.

[95] Gilles Vandewiele, Femke Ongenae, and Filip De Turck. Gendis: Genetic discovery of shapelets. Sensors, 21(4):1059, Feb 2021.

[96] Roberto Medico, Joeri Ruyssinck, Dirk Deschrijver, and Tom Dhaene. Learning multivariate shapelets with multi-layer neural networks for interpretable time-series classification. Advances in Data Analysis and Classification, 15, 03 2021.

[97] Riccardo Guidotti and Matteo D'Onofrio. Matrix Profile-Based Interpretable Time Series Classifier. Frontiers Artif. Intell., 4:699448, 2021.

[98] Riccardo Guidotti and Anna Monreale. Designing Shapelets for Interpretable Data-Agnostic Classification. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, page 532–542, New York, NY, USA, 2021. Association for Computing Machinery.

[99] Yupeng Hu, Peng Zhan, Yang Xu, Jia Zhao, Yujun Li, and Xueqing Li. Temporal representation learning for time series classification. Neural Computing and Applications, 33(8):3169–3182, 2021.

[100] Riccardo Guidotti, Anna Monreale, Francesco Spinnato, Dino Pedreschi, and Fosca Giannotti. Explaining Any Time Series Classifier. In 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI), pages 167–176, 2020.

[101] Dominique Mercier, Andreas Dengel, and Sheraz Ahmed. PatchX: Explaining Deep Models by Intelligible Pattern Patches for Time-series Classification. In 2021 International Joint Conference on Neural Networks (IJCNN), pages 1–8, 2021.

[102] Dominique Mercier, Andreas Dengel, and Sheraz Ahmed. P2ExNet: Patch-based prototype explanation network. In International Conference on Neural Information Processing, pages 318–330. Springer, 2020.

[103] Subhajit Das, Panpan Xu, Zeng Dai, Alex Endert, and Liu Ren. Interpreting Deep Neural Networks through Prototype Factorization. In 2020 International Conference on Data Mining Workshops (ICDMW), pages 448–457, 2020.

[104] Jingyuan Wang, Ze Wang, Jianfeng Li, and Junjie Wu. Multilevel wavelet decomposition network for interpretable time series analysis. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2437–2446, 2018.

[105] Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. Interpretable and Steerable Sequence Learning via Prototypes. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Jul 2019.

[106] Wensi Tang, Lu Liu, and Guodong Long. Interpretable Time-series Classification on Few-shot Samples. In 2020 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE, 2020.

[107] Xuchao Zhang, Yifeng Gao, Jessica Lin, and Chang-Tien Lu. Tapnet: Multivariate time series classification with attentional prototypical network. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 6845–6852, 2020.

[108] Dominique Gay, Romain Guigourès, Marc Boullé, and Fabrice Clérot. Feature extraction over multiple representations for time series classification. In International Workshop on New Frontiers in Mining Complex Patterns, pages 18–34. Springer, 2013.

[109] Vera Shalaeva, Sami Alkhoury, Julien Marinescu, Cécile Amblard, and Gilles Bisson. Multi-operator Decision Trees for Explainable Time-Series Classification, pages 86–99. Springer International Publishing, 01 2018.

[110] Hidetoshi Ito and Basabi Chakraborty. A proposal for shape aware feature extraction for time series classification. In iCAST, pages 1–6. IEEE, 2019.

[111] Ferdinand Küsters, Peter Schichtel, Sheraz Ahmed, and Andreas Dengel. Conceptual explanations of neural network prediction for time series. In 2020 International Joint Conference on Neural Networks (IJCNN), pages 1–6. IEEE, 2020.

[112] Munawar Zaman and Adnan Hassan. Fuzzy Heuristics and Decision Tree for Classification of Statistical Feature-Based Control Chart Patterns. Symmetry, 13(1), 2021.

[113] Eoin Delaney, Derek Greene, and Mark T. Keane. Instance-based counterfactual explanations for time series classification. In Antonio A. Sánchez-Ruiz and Michael W. Floyd, editors, Case-Based Reasoning Research and Development, pages 32–47, Cham, 2021. Springer International Publishing.

[114] Isak Karlsson, Jonathan Rebane, Panagiotis Papapetrou, and Aristides Gionis. Locally and globally explainable time series tweaking. Knowledge and Information Systems, 62, 05 2020.

[115] Emre Ates, Burak Aksar, Vitus J. Leung, and Ayse K. Coskun. Counterfactual Explanations for Multivariate Time Series. 2021 International Conference on Applied Artificial Intelligence (ICAPAI), May 2021.

[116] Jokin Labaien, Ekhi Zugasti, and Xabier De Carlos. Contrastive Explanations for a Deep Learning Model on Time-Series Data. In DaWaK, 2020.

[117] Yuzuru Okajima and Kunihiko Sadamasa. Deep neural networks constrained by decision rules. In AAAI, pages 2496–2505. AAAI Press, 2019.

[118] Hao Huang, Chenxiao Xu, Shinjae Yoo, Weizhong Yan, Tianyi Wang, and Feng Xue. Imbalanced Time Series Classification for Flight Data Analyzing with Nonlinear Granger Causality Learning. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pages 2533–2540, 2020.

[119] Sara Mohammadinejad, Jyotirmoy V Deshmukh, Aniruddh G Puranic, Marcell Vazquez-Chanlatte, and Alexandre Donzé. Interpretable classification of time-series data using efficient enumerative techniques. In Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control, pages 1–10, 2020.

[120] Udo Schlegel and Daniel A Keim. Time series model attribution visualizations as explanations. In 2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX), pages 27–31. IEEE, 2021.

[121] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings, 2014.

[122] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods. Advances in Neural Information Processing Systems, 33:4211–4222, 2020.

[123] David S Watson. Conceptual challenges for interpretable machine learning. Synthese, 200(1):1–33, 2022.

[124] Daniel Fryer, Inga Strümke, and Hien Nguyen. Shapley values for feature selection: the good, the bad, and the axioms. IEEE Access, 9:144352–144360, 2021.

[125] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 27, 2014.

[126] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems, 30, 2017.

[127] Sarthak Jain and Byron C Wallace. Attention is not explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3543–3556, 2019.

[128] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing SAX: A Novel Symbolic Representation of Time Series. Data Min. Knowl. Discov., 15:107–144, 08 2007.

[129] Eamonn J. Keogh and Michael J. Pazzani. Scaling up Dynamic Time Warping for Datamining Applications. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '00, page 285–289, New York, NY, USA, 2000. Association for Computing Machinery.

[130] Patrick Schäfer and Mikael Högqvist. SFA: A symbolic fourier approximation and index for similarity search in high dimensional datasets. ACM International Conference Proceeding Series, pages 516 – 527, 2012.

[131] Lexiang Ye and Eamonn Keogh. Time series shapelets: A new primitive for data mining. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, page 947–956, New York, NY, USA, 2009.

[132] Thanawin Rakthanmanon and Eamonn Keogh. Fast Shapelets: A Scalable Algorithm for Discovering Time Series Shapelets, pages 668–676. Society for Industrial and Applied Mathematics, 2013.

[133] Aaron Bostrom and Anthony Bagnall. Binary shapelet transform for multiclass time series classification. In Sanjay Madria and Takahiro Hara, editors, Big Data Analytics and Knowledge Discovery, pages 257–269, Cham, 2015. Springer International Publishing.

[134] Xavier Renard, Maria Rifqi, and Marcin Detyniecki. Random-shapelet: An algorithm for fast shapelet discovery. In 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 10 2015.

[135] Lu Hou, James T. Kwok, and Jacek M. Zurada. Efficient learning of timeseries shapelets. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, page 1209–1215. AAAI Press, 2016.

[136] Cun Ji, Chao Zhao, Shijun Liu, Chenglei Yang, Li Pan, Lei Wu, and Xiangxu Meng. A fast shapelet selection algorithm for time series classification. Computer Networks, 148:231–240, 2019.

[137] Jingwei Zuo, Karine Zeitouni, and Yehia Taher. Exploring Interpretable Features for Large Time Series with SE4TeC. In EDBT, 03 2019.

[138] Chao Zhao, Shijun Liu, Li Pan, Cun Ji, and Chenglei Yang. Selecting superior candidates from a suitable set: A selective extraction algorithm for accelerating shapelet discovery in time series data. In 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD), pages 404–409. IEEE, 2019.

[139] Guozhong Li, Byron Koon Kau Choi, Jianliang Xu, Sourav S Bhowmick, Kwok-Pan Chun, and Grace LH Wong. Efficient shapelet discovery for time series classification. IEEE Transactions on Knowledge and Data Engineering, 2020.

[140] Om P Patri, Anand V Panangadan, Charalampos Chelmis, and Viktor K Prasanna. Extracting discriminative features for event-based electricity disaggregation. In 2014 IEEE Conference on Technologies for Sustainability (SusTech), pages 232–238. IEEE, 2014.

[141] Mustafa Cetin, Abdullah Mueen, and Vince Calhoun. Shapelet Ensemble for Multi-dimensional Time Series. In SDM, 04 2015.

[142] Josif Grabocka, Martin Wistuba, and Lars Schmidt-Thieme. Fast classification of univariate and multivariate time series through shapelet discovery. Knowledge and Information Systems, 49, 11 2016.

[143] Mit Shah, Josif Grabocka, Nicolas Schilling, Martin Wistuba, and Lars Schmidt-Thieme. Learning DTW-Shapelets for Time-Series Classification. In Proceedings of the 3rd IKDD Conference on Data Science, 2016, CODS '16, New York, NY, USA, 2016. Association for Computing Machinery.

[144] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn J. Keogh. Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. In ICDM, pages 1317–1322. IEEE Computer Society, 2016.

[145] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. Journal of machine learning research, 9(11), 2008.

[146] Isak Karlsson, Panagiotis Papapetrou, and Henrik Boström. Generalized random shapelet forests. Data Min. Knowl. Discov., 30(5):1053–1085, sep 2016.

[147] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. Advances in Neural Information Processing Systems, 31, 2018.

[148] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech., 31:841, 2017.

[149] Argyro Kampouraki, George Manis, and Christophoros Nikou. Heartbeat time series classification with support vector machines. IEEE Transactions on Information Technology in Biomedicine, 13(4):512–518, 2008.

[150] Wanpracha Art Chaovalitwongse, Oleg A Prokopyev, and Panos M Pardalos. Electroencephalogram (EEG) time series classification: Applications in epilepsy. Annals of Operations Research, 148(1):227–250, 2006.

[151] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In Twenty-fourth international joint conference on artificial intelligence, 2015.

[152] Pierre Geurts. Pattern extraction for time series classification. In European conference on principles of data mining and knowledge discovery, pages 115–127. Springer, 2001.

[153] Pei Li, Mohamed Abdel-Aty, and Jinghui Yuan. Real-time crash risk prediction on arterials based on LSTM-CNN. Accident Analysis & Prevention, 135:105371, 2020.

[154] Gilseung Ahn, Hwanchul Lee, Jisu Park, and Sun Hur. Development of indicator of data sufficiency for feature-based early time series classification with applications of bearing fault diagnosis. Processes, 8(7):790, 2020.

[155] Min Su Kim, Jong Pil Yun, and PooGyeon Park. An Explainable Convolutional Neural Network for Fault Diagnosis in Linear Motion Guide. IEEE Transactions on Industrial Informatics, page 1, 2020.

[156] Xiaoyin Nie and Gang Xie. A novel normalized recurrent neural network for fault diagnosis with noisy labels. Journal of Intelligent Manufacturing, 2020.

[157] Praharsh Ivaturi, Matteo Gadaleta, Amitabh C Pandey, Michael Pazzani, Steven R Steinhubl, and Giorgio Quer. A comprehensive explanation framework for biomedical time series classification. IEEE Journal of Biomedical and Health Informatics, 25(7):2398–2408, 2021.

[158] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. International Journal of Computer Assisted Radiology and Surgery, 14(9):1611–1617, Jul 2019.

[159] Xiang Zhang, Lina Yao, Manqing Dong, Zhe Liu, Yu Zhang, and Yong Li. Adversarial representation learning for robust patient-independent epileptic seizure detection. IEEE J. Biomed. Health Inform., 24(10):2852–2859, 2020.

[160] Theekshana Dissanayake, Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for patient-independent epileptic seizure prediction using scalp eeg signals. IEEE Sensors Journal, 21(7):9377–9388, 2021.

[161] Mennatallah El-Assady, Wolfgang Jentner, Rebecca Kehlbeck, Udo Schlegel, Rita Sevastjanova, Fabian Sperrle, Thilo Spinner, and Daniel Keim. Towards XAI: Structuring the Processes of Explanations. In ACM Workshop on Human-Centered Machine Learning, 2019.

[162] Dominique Mercier, Jwalin Bhatt, Andreas Dengel, and Sheraz Ahmed. Time to Focus: A Comprehensive Benchmark Using Time Series Attribution Methods. arXiv preprint arXiv:2202.03759, 2022.

[163] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31, 2018.

[164] Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. Advances in Neural Information Processing Systems, 33:6441–6452, 2020.

[165] Thu Trang Nguyen, Thach Le Nguyen, and Georgiana Ifrim. A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification. In International Workshop on Advanced Analytics and Learning on Temporal Data, pages 77–94. Springer, 2020.

[166] Udo Schlegel, Daniela Oelke, Daniel A. Keim, and Mennatallah El-Assady. An Empirical Study of Explainable AI Techniques on Deep Learning Models For Time Series Tasks. NeurIPS 2020 Workshops, 2020.

[167] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. Advances in Neural Information Processing Systems, 32, 2019.

[168] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. Advances in Neural Information Processing Systems, 32, 2019.

[169] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch, 2020.

[170] Jacob Gildenblat and contributors. PyTorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

[171] Markus Löning, A. Bagnall, Sajaysurya Ganesh, Viktor Kazakov, Jason Lines, and Franz J. Király. sktime: A unified interface for machine learning with time series. arXiv, 2019.

[172] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. Tslearn, a machine learning toolkit for time series data. Journal of Machine Learning Research, 21(118):1–6, 2020.

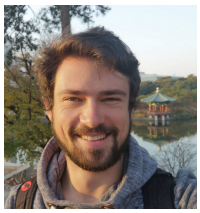
[173] Johann Faouzi and Hicham Janati. pyts: A python package for time series classification. Journal of Machine Learning Research, 21(46):1–6, 2020.

[174] Rita Sevastjanova, Fabian Beck, Basil Ell, Cagatay Turkay, Rafael Henkin, Miriam Butt, Daniel Keim, and Mennatallah El-Assady. Going beyond visualization. verbalization as complementary medium to explain machine learning models. Workshop on Visualization for AI Explainability at IEEE VIS, 2018.

[175] Q Vera Liao and Kush R Varshney. Human-centered explainable AI (XAI): From algorithms to user experiences. arXiv preprint arXiv:2110.10790, 2021.
[176] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923, 2017.
[177] Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In Proceedings of the 2020 conference on fairness, accountability, and transparency, pages 629–639, 2020.
[178] Jose N Paredes, Juan Carlos L Teze, Gerardo I Simari, and Maria Vanina Martinez. On the Importance of Domain-specific Explanations in AI-based Cybersecurity Systems (Technical Report). arXiv preprint arXiv:2108.02006, 2021.
[179] Hoang Anh Dau, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, Yanping, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, Gustavo Batista, and Hexagon-ML. The UCR Time Series Classification Archive, October 2018. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.
[180] Mohamed Karim Belaid, Eyke Hüllermeier, Maximilian Rabus, and Ralf Krestel. Do We Need Another Explainable AI Method? Toward Unifying Post-hoc XAI Evaluation Methods into an Interactive and Multi-dimensional Benchmark. arXiv preprint arXiv:2207.14160, 2022.
[181] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
[182] Leila Arras, Ahmed Osman, and Wojciech Samek. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion, 81:14–40, 2022.

ANDREAS THEISSLER is a professor at Aalen University of Applied Sciences, Germany, where he researches and lectures on different aspects of Machine Learning and Human-Centered AI. He received his PhD from Brunel University London in 2014, prior to that he studied Software Engineering and has worked in different Data Science positions in industry. He has published on the interplay of Machine Learning and users, for example on the questions of how we can evaluate, understand, improve, or enable Machine Learning by incorporating expert knowledge. In addition, he has worked in the field of machine learning in applications, e.g. anomaly detection in time series from automotive systems.

FRANCESCO SPINNATO is a Ph.D. student in Data Science at Scuola Normale Superiore in Pisa, Italy. His research focus is on explainable AI for sequential data, in particular on interpreting black-box models for univariate and multivariate time series. In 2017, he received his B.S. degree in Economics and Management at the University of Padua, Italy, and in 2020 his M.S. degree in Data Science at the University of Pisa, Italy.



UDO SCHLEGEL received a B.S. degree in Information Engineering in 2016 and an M.S. degree in Computer and Information Science in 2018 from the University of Konstanz. He is pursuing a Ph.D. in Computer Science at the University of Konstanz. His research interest includes developing and evaluating explainable AI techniques for deep learning models in the time series domain with a focus on attribution approaches. Further research interests include graph neural networks, visual analytics for deep learning, and deep learning applications in the wild.

RICCARDO GUIDOTTI is an Assistant Professor (RTD-B) at the Department of Computer Science University of Pisa, Italy and a member of the Knowledge Discovery and Data Mining Laboratory (KDDLab), a joint research group with the Information Science and Technology Institute of the National Research Council in Pisa. In 2013 and 2010 he graduated cum laude in Computer Science (MS and BS) at University of Pisa. He received the PhD in Computer Science with a thesis on Personal Data Analytics in the same institution. His research interests are in XAI, interpretable ML, Personal Data Analytics, and Quantum Computing.