A Performance-Explainability Framework to Benchmark Machine Learning Methods: Application to Multivariate Time Series Classifiers

Kevin Fauvel, Véronique Masson, Elisa Fromont

A Performance-Explainability Framework to Benchmark Machine Learning Methods: Application to Multivariate Time Series Classifiers

Kevin Fauvel , **Véronique Masson** and **Élisa Fromont**
Univ Rennes, Inria, CNRS, IRISA, France
kevin.fauvel@inria.fr, {veronique.masson, elisa.fromont}@irisa.fr

Abstract

Our research aims to propose a new performance-explainability analytical framework to assess and benchmark machine learning methods. The framework details a set of characteristics that systematize the performance-explainability assessment of existing machine learning methods. In order to illustrate the use of the framework, we apply it to benchmark the current state-of-the-art multivariate time series classifiers.

1 Introduction

There has been an increasing trend in recent years to leverage machine learning methods to automate decision-making processes. However, for many applications, the adoption of such methods cannot rely solely on their prediction performance. For example, the European Union's General Data Protection Regulation, which became enforceable on 25 May 2018, introduces a right to explanation for all individuals so that they can obtain "meaningful explanations of the logic involved" when automated decision-making has "legal effects" on individuals or similarly "significantly affecting" them[1]. Therefore, in addition to their prediction performance, machine learning methods have to be assessed on how they can supply their decisions with explanations.

The performance of a machine learning method can be assessed by the extent to which it correctly predicts unseen instances. A metric like the accuracy score commonly measures the performance of a classification model. However, there is no standard approach to assess explainability. First, there is no mathematical definition of explainability. A definition proposed by [Miller, 2019] states that the higher the explainability of a machine learning algorithm, the easier it is for someone to comprehend why certain decisions or predictions have been made. Second, there are several methods belonging to different categories (explainability-by-design, post-hoc model-specific explainability and post-hoc model-agnostic explainability) [Du *et al.*, 2020], which provide their own form of explanations.

The requirements for explainable machine learning methods are dependent upon the application and to whom the explanations are intended for [Tomsett *et al.*, 2018; Bohlender and Kóhl, 2019]. In order to match these requirements and conduct experiments to validate the usefulness of the explanations by the end-users, there is a need to have a comprehensive assessment of the explainability of the existing methods. Doshi-Velez and Kim [2017] claim that creating a shared language is essential for the evaluation and comparison of machine learning methods, which is currently challenging without a set of explanation characteristics. As far as we have seen, there is no existing framework which defines a set of explanation characteristics that systematize the assessment of the explainability of existing machine learning methods.

Hence, in this paper, we propose a new framework to assess and benchmark the performance-explainability characteristics of machine learning methods. The framework hypothesizes a set of explanation characteristics, and as emphasized in [Wolf, 2019], focuses on what people might need to understand about machine learning methods in order to act in concert with the model outputs. The framework does not claim to be exhaustive and excludes application-specific implementation constraints like time, memory usage and privacy. It could be a basis for the development of a comprehensive assessment of the machine learning methods with regards to their performance and explainability and for the design of new machine learning methods. Due to space constraint, we limit the illustration of the use of the framework to one category of machine learning methods and we choose the Multivariate Time Series (MTS) classifiers. Multivariate data which integrates temporal evolution has received significant interests over the past decade, driven by automatic and high-resolution monitoring applications (e.g. healthcare [Li *et al.*, 2018], mobility [Jiang *et al.*, 2019], natural disasters [Fauvel *et al.*, 2020a]). Moreover, the available explainability solutions to support the current state-of-the-art MTS classifiers remain limited, so this category of methods appears meaningful to assess for us.

The contributions of this paper are the following:

- We present a new performance-explainability analytical framework to assess and benchmark machine learning methods;
- We detail a set of characteristics that systematize the performance-explainability assessment of existing machine learning methods;
- We illustrate the use of the framework by benchmarking the current state-of-the-art MTS classifiers.

[1] https://ec.europa.eu/info/law/law-topic/data-protection_en

2 Related Work

In this section, we first position this paper in the related work and introduce the different categories of explainability methods as a background to the notions that will be discussed in the framework. Then, we present the state-of-the-art machine learning methods that will be used to illustrate the framework, i.e. MTS classifiers.

2.1 Explainability

Multiple taxonomies of explainability methods have been derived from different frameworks [Guidotti *et al.*, 2018; Ventocilla *et al.*, 2018; Du *et al.*, 2020]. However, none of them defines a set of explanation characteristics that systematize the assessment of the explainability of existing machine learning methods. [Guidotti *et al.*, 2018] provides a classification of the main problems addressed in the literature with respect to the notion of explanation and the type of machine learning systems. [Ventocilla *et al.*, 2018] proposes a high-level taxonomy of interpretable and interactive machine learning composed of six elements (Dataset, Optimizer, Model, Predictions, Evaluator and Goodness). And, [Du *et al.*, 2020] categorizes existing explainability methods of machine learning models into either by design or post-hoc explainability. As our framework aims to cover all types of methods, we do not present the frameworks focusing on a particular type of explainability methods (e.g. [Lundberg and Lee, 2017; Ancona *et al.*, 2018; Henin and Métayer, 2019]).

A five-step method to understand the requirements for explainable AI systems has been published in [Hall *et al.*, 2019]. The five steps are: explainee role definition, explanation characteristics identification, requirements collection, existing methods assessment and requirements/existing methods mapping. Our framework can be positioned as a further development of the fourth step of the method by detailing a set of explanations characteristics that systematize the assessment of existing methods. Our framework does not include application-specific implementation constraints like time, memory usage and privacy.

As a background to the notions that will be discussed in the framework, we introduce the three commonly recognized categories (explainability-by-design, post-hoc model-specific explainability and post-hoc model-agnostic explainability) [Du *et al.*, 2020] to which all of the explainability methods are belonging to. First, some machine learning models provide explainability-by-design. These self-explanatory models incorporate explainability directly to their structures. This category includes, for example, decision trees, rule-based models and linear models. Next, post-hoc model-specific explainability methods are specifically designed to extract explanations for a particular model. These methods usually derive explanations by examining internal model structures and parameters. For example, a method has been designed to measure the contribution of each feature in random forests [Palczewska *et al.*, 2013]; and another one has been designed to identify the regions of input data that are important for predictions in convolutional neural networks using the class-specific gradient information [Selvaraju *et al.*, 2019]. Finally, post-hoc model-agnostic explainability methods provide explanations from any machine learning model. These methods treat the model as a black-box and does not inspect internal model parameters. For example, the permutation feature importance method [Altmann *et al.*, 2010] and the methods using an explainable surrogate model [Lakkaraju *et al.*, 2017; Lundberg and Lee, 2017; Ribeiro *et al.*, 2018; Guidotti *et al.*, 2019] belong to this category.

The explainability methods presented reflect the diversity of explanations generated to support model predictions, therefore the need for a framework in order to benchmark the machine learning methods explainability. The next section present the MTS classifiers that will be used to illustrate the framework.

2.2 Multivariate Time Series Classifiers

The state-of-the-art MTS classifiers consist of a diverse range of methods which can be categorized into three families: similarity-based, feature-based and deep learning methods.

Similarity-based methods make use of similarity measures to compare two MTS. Dynamic Time Warping (DTW) has been shown to be the best similarity measure to use along the k-Nearest Neighbors (k-NN) [Seto *et al.*, 2015]. There are two versions of kNN-DTW for MTS: dependent (DTW_D) and independent (DTW_I). Neither dominates over the other [Shokoohi-Yekta *et al.*, 2017] from an accuracy perspective but DTW_I allows the analysis of distance differences at feature level.

Next, feature-based methods include shapelets (gRSF [Karlsson *et al.*, 2016], UFS [Wistuba *et al.*, 2015]) and bag-of-words (LPS [Baydogan and Runger, 2016], mv-ARF [Tuncel and Baydogan, 2018], SMTS [Baydogan and Runger, 2014], WEASEL+MUSE [Schäfer and Leser, 2017]) models. WEASEL+MUSE shows better results compared to gRSF, LPS, mv-ARF, SMTS and UFS on average (20 MTS datasets). WEASEL+MUSE generates a bag-of-words representation by applying various sliding windows with different sizes on each discretized dimension (Symbolic Fourier Approximation) to capture features (unigrams, bigrams, dimension identification). Following a feature selection with chi-square test, it classifies the MTS based on a logistic regression classifier.

Then, deep learning methods use Long-Short Term Memory (LSTM) and/or Convolutional Neural Networks (CNN). According to the results published, the current state-of-the-art model (MLSTM-FCN) is proposed in [Karim *et al.*, 2019] and consists of a LSTM layer and a stacked CNN layer along with Squeeze-and-Excitation blocks to generate latent features.

Therefore, we choose to benchmark the performance-explainability of the best-in-class for each similarity-based, feature-based and deep learning category (DTW_I, WEASEL+MUSE and MLSTM-FCN classifiers). The next section introduces the performance-explainability framework, which is illustrated with the benchmark of the best-in-class MTS classifiers in section 4.

3 Performance-Explainability Framework

The framework aims to respond to the different questions an end-user may ask to take an informed decision based on the predictions made by a machine learning model: *What is the*

level of performance of the model? Is the model comprehensible? Is it possible to get an explanation for a particular instance? Which kind of information does the explanation provide? Can we trust the explanations? What is the target user category of the explanations? The performance-explainability framework that we propose is composed of the following components, which will also be translated into terms specific to our application (MTS classifiers) whenever relevant:

Performance *What is the level of performance of the model?* The first component of the framework characterizes the performance of a machine learning model. Different methods (e.g. holdout, k-fold cross-validation) and metrics (e.g. accuracy, F-measure, Area Under the ROC Curve) exist to evaluate the performance of a machine learning model [Witten *et al.*, 2016]. However, there is no consensus on an evaluation procedure to assess the performance of a machine learning model. Recent work suggests that the definition of such an evaluation procedure necessitates the development of a measurement theory for machine learning [Flach, 2019]. Many of the problems stem from a limited appreciation of the importance of the *scale* on which the evaluation measures are expressed.

Then, in current practices, the choice of a metric to evaluate the performance of a machine learning model depends on the application. According to the application, a metric aligned with the goal of the experiments is selected, which prevents the performance comparison of machine learning models across applications.

Therefore, the performance component in the framework is defined as a first step towards a standard procedure to assess the performance of machine learning models. It corresponds to the relative performance of a model on a particular application. More specifically, it indicates the relative performance of the models as compared to the state-of-the-art model on a particular application and an evaluation setting. This definition allows the categorization of the models' performance on an application and an evaluation setting. In the case of different applications with a similar machine learning task, the performance component can give the list of models which outperformed current state-of-the-art models on their respective application. Thus, it points to certain models that could be interesting to evaluate on a new application, without providing guarantee that these models would perform the same on this new application. We propose an assessment of the performance in three categories:

- *Best*: best performance. It corresponds to the performance of the first ranked model on the application following an evaluation setting (models, evaluation method, datasets);
- *Similar*: performance similar to that of the state-of-the-art models. Based on the same evaluation setting, it corresponds to all the models which do not show a statistically significant performance difference with the second ranked model. For example, the statistical comparison of multiple classifiers on multiple datasets is usually presented on a critical difference diagram [Demšar, 2006];
- *Below*: performance below that of the state-of-the-art models. It corresponds to the performance of the remaining models with the same evaluation setting.

Model Comprehensibility *Is the model comprehensible?* The model comprehensibility corresponds to the ability for the user to understand how the model works and produces certain predictions. Comprehensibility is tightly linked to the model complexity; yet, there is no consensus on model complexity assessment [Guidotti *et al.*, 2018]. Currently, two categories of models are commonly recognized: "white-box" models, i.e. easy-to-understand models, and "black-box" models, i.e. complicated-to-understand models [Lipton, 2016]. For example, many rule-based models and decision trees are regarded as "white-box" models while ensemble methods and deep learning models are "black-box" models. Not all rule-based models or decision trees are "white-box" models. Cognitive limitations of humans place restrictions on the complexity of the approximations that are understandable to humans. For example, a decision tree with a hundred levels cannot be considered as an easy-to-understand model [Lakkaraju *et al.*, 2017].

Nevertheless, the distinction between "white-box" models and "black-box" models is clear among the machine learning methods of this paper. The state-of-the-art MTS classifiers are all "black-box" except one which is an easy-to-understand similarity-based approach. Therefore, due to space limitation, we propose a first assessment of the comprehensibility in two categories and we plan to further elaborate this component in future work:

- *Black-Box*: "black-box" model, i.e. complicated-to-understand models;
- *White-Box*: "white-box" model, i.e. easy-to-understand models.

Granularity of the Explanations *Is it possible to get an explanation for a particular instance?* The granularity indicates the level of possible explanations. Two levels are generally distinguished: global and local [Du *et al.*, 2020]. Global explainability means that explanations concern the overall behavior of the model across the full dataset, while local explainability informs the user about a particular prediction. Some methods can provide either global or local-only explainability while other methods can provide both (e.g. decision trees). Therefore, we propose an assessment of the granularity in three categories:

- *Global*: global explainability;
- *Local*: local explainability;
- *Global & Local*: both global and local explainability.

Information Type *Which kind of information does the explanation provide?* The information type informs the user about the kind of information communicated. The most valuable information is close to the language of human reasoning, with causal and counterfactual rules [Pearl and Mackenzie, 2018]. Causal rules can tell the user that certain observed variables are the causes of specific model predictions. However, machine learning usually leverages statistical associations in the data and do not convey information about the causal relationships among the observed variables and the unobserved confounding variables. The usual statistical asso-

ciations discovered by machine learning methods highly depend on the machine learning task. Therefore, we first give a generic high-level definition of the information type and then we detail and illustrate it for the application case of this paper (MTS classification). We propose a generic assessment of the information type in 3 categories from the least to the most informative:

- *Importance*: the explanations reveal the relative importance of each dataset variable on predictions. The importance indicates the statistical contribution of each variable to the underlying model when making decisions;
- *Patterns*: the explanations provide the small conjunctions of symbols with a predefined semantic (patterns) associated with the predictions;
- *Causal*: the most informative category corresponds to explanations under the form of causal rules;

In this paper, the issue of Multivariate Time Series (MTS) classification is addressed. A MTS $M = \{x_1, ..., x_d\} \in \mathcal{R}^{d*l}$ is an ordered sequence of $d \in \mathcal{N}$ streams with $x_i = (x_{i,1}, ..., x_{i,l})$, where l is the length of the time series and d is the number of multivariate dimensions. Thus, considering the MTS data type, the information can be structured around the features, i.e. the observed variables, and the time. We propose to decompose the 3 categories presented into 8 categories. In addition, we will illustrate each of these categories with an application in the medical field. Figure 1 shows the first MTS of the UEA Atrial Fibrilation [Bagnall *et al.*, 2018] test set that belongs to the class *Non-Terminating Atrial Fibrilation*. This MTS is composed of two dimensions (two channels ECG) with a length of 640 (5 second period with 128 samples per second). It is worth noting that the explanations provided to illustrate each category are assumptive rather than validated, they are given as illustrative in nature.

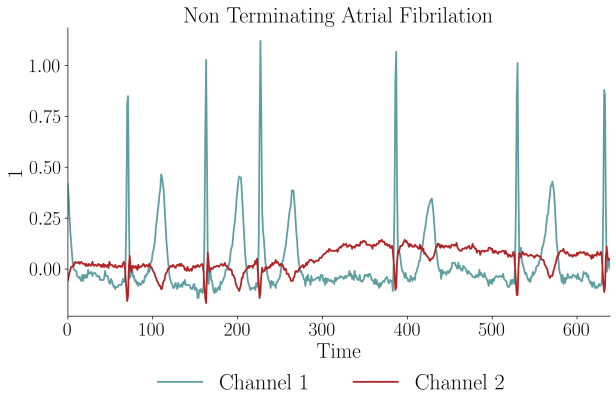

Figure 1: The first MTS sample of the UEA Atrial Fibrilation test set. It belongs to the class *Non-Terminating Atrial Fibrilation* and is composed of two channels ECG on a 5 second period (128 samples per second).

- *Features* (*Importance*): the explanations reveal the relative importance of the features on predictions. For example, in order to support a model output from the MTS of the Figure 1, the explanations could tell the user that the channel 2 has a greater importance on the prediction than the channel 1;
- *Features* + *Time* (*Importance*): the explanations provide the relative importance of the features and timestamps on predictions. For example, in order to support a model output from the MTS of the Figure 1, the explanations could tell the user that the channel 2 has a greater importance on the prediction than the channel 1 and that the timestamps are in increasing order of importance on the prediction;
- *Features* + *Time* + *Values* (*Importance*): in addition to the relative importance of the features and timestamps on predictions, the explanations indicate the discriminative values of a feature for each class. For example in Figure 1, the explanations could give the same explanations as the previous category, plus, it could tell the user that the timestamps with the highest importance are associated with high values (values above 0.15) on the channel 2;
- *Uni Itemsets* (*Patterns*): the explanations provide patterns under the form of groups of values, also called itemsets, which occur per feature and are associated with the prediction. For example, in order to support a model output from the MTS of the Figure 1, the explanations could tell the user that the following itemsets are associated with the prediction: {channel 1: extremely high value (above 1); channel 1: low value (below -0.05)} and {channel 2: high value (above 0.15); channel 2: extremely low value (below -0.1)}. The first itemset can be read as: the prediction is associated with the occurence on the channel 1 of an extremely high value being above 1 and a low value being below -0.05 at another moment, without information on which one appears first;
- *Multi Itemsets* (*Patterns*): the explanations provide patterns under the form of multidimensional itemsets, i.e. groups of values composed of different features, which are associated with the prediction. For example, in order to support a model output from the MTS of the Figure 1, the explanations could tell the user that the following itemset is associated with the prediction: {channel 1: extremely high value (above 1); channel 2: high value (above 0.15)};
- *Uni Sequences* (*Patterns*): the explanations provide patterns under the form of ordered groups of values, also called sequences, which occur per feature and are associated with the prediction. For example, in order to support a model output from the MTS of the Figure 1, the explanations could tell the user that the following sequences are associated with the prediction: <channel 1: extremely high value (above 1); channel 1: low value (below -0.05)> and <channel 2: high values (above 0.15) with an increase during 1 second>. The first sequence can be read as: the prediction is associated with the occurrence on the channel 1 of an extremely high value being above 1 followed by a low value being below -0.05;
- *Multi Sequences* (*Patterns*): the explanations provide patterns under the form of multidimensional sequences, i.e. ordered groups of values composed of different features, which are associated with the prediction. For example, in order to support a model output from the MTS

Table 1: Summary of framework results of the state-of-the-art MTS classifiers.

	Similarity-Based DTW_I	**Feature-Based** WEASEL+MUSE with SHAP	**Deep Learning** MLSTM-FCN with SHAP
Performance	Below[1]	Similar[1]	Best[1]
Comprehensibility	White-Box	Black-Box	Black-Box
Granularity	Local	Both Global & Local	Both Global & Local
Information	Features+Time	Features+Time	Features+Time
Faithfulness	Perfect	Imperfect	Imperfect
User	Domain Expert	Domain Expert	Domain Expert

[1] Predefined train/test splits and an arithmetic mean of the accuracies on 35 public datasets [Karim et al., 2019]. As presented in section 2.2, the models evaluated in the benchmark are: DTW_D, DTW_I, gRSF, LPS, MLSTM-FCN, mv-ARF, SMTS, UFS and WEASEL+MUSE.

of the Figure 1, the explanations could tell the user that the following sequence is associated with the prediction: <channel 1: extremely high value (above 1); channel 2: high values (above 0.15) with an increase during 1 second>;

- *Causal*: the last category corresponds to explanations under the form of causal rules. For example, in order to support a model output from the MTS of the Figure 1, the explanations could tell the user that the following rule applies: if (channel 1: extremely high value (above 1)) & (channel 2: high values (above 0.15) with an increase during 1 second), then the MTS belongs to the class *Non-Terminating Atrial Fibrilation*.

Faithfulness *Can we trust the explanations?* The faithfulness corresponds to the level of trust an end-user can have in the explanations of model predictions, i.e. the level of relatedness of the explanations to what the model actually computes. An explanation extracted directly from the original model is faithful by definition. Some post-hoc explanation methods propose to approximate the behavior of the original "black-box" model with an explainable surrogate model. The explanations from the surrogate models cannot be perfectly faithful with respect to the original model [Rudin, 2019]. The fidelity criteria is used to quantify the faithfulness by the extent to which the surrogate model imitates the prediction score of the original model [Guidotti *et al.*, 2018].

In this paper, two MTS classifiers use an explainable surrogate model among the three state-of-the-art methods presented in section 4. However, there is no need to distinguish between the degree of fidelity of the surrogate models for the purpose of the comparison in this paper. Therefore, due to space limitation, we propose a first assessment of the faithfulness in two categories and we plan to further elaborate this component in future work:

- *Imperfect*: imperfect faithfulness (use of an explainable surrogate model);
- *Perfect*: perfect faithfulness.

User category *What is the target user category of the explanations?* The user category indicates the audience to whom the explanations are accessible. The user's experience will affect what kind of *cognitive chunks* they have, that is, how they organize individual elements of information into collections [Neath and Surprenant, 2003]. Thus, it could be interesting to categorize the user types and associate with the model to whom the explanations will be accessible to. The broader the audience, the better are the explanations. Therefore, we propose an assessment in three categories:

- *Machine Learning Expert*;
- *Domain Expert*: domain experts (e.g. professionals, researchers);
- *Broad Audience*: non-domain experts (e.g. policy makers).

In order to compare the methods visually using the proposed framework, the different aspects can be represented on a parallel coordinates plot. A parallel coordinate plot allows a 2-dimensional visualization of a high dimensional dataset and is suited for the categorical data of this framework. The next section presents an example of parallel coordinates plots comparing the state-of-the-art MTS classifiers.

4 Application to Multivariate Time Series Classifiers

This section shows how the framework presented in the previous section can be used to assess and benchmark the state-of-the-art MTS classifiers. As introduced in section 2.2, the state-of-the-art MTS classifiers are: DTW_I, MLSTM-FCN and WEASEL+MUSE. The results of the assessment are summarized in Table 1, illustrated in Figure 2 and detailed in the following paragraphs.

The first MTS classifier belongs to the similarity-based category and is the one-nearest neighbor MTS classifier with DTW distance (DTW_I). DTW_I classifies MTS based on the label of the nearest sample and a similarity calculated as the cumulative distances of all dimensions independently measured under DTW. For each MTS, the explanation supporting the classification is the ranking of features and timestamps in decreasing order of their DTW distance with the nearest MTS. Based on predefined train/test splits and an arithmetic mean of the accuracies, DTW_I underperforms the current state-of-the-art MTS classifiers on the 35 public datasets (Performance: *Below*). The results from [Karim *et al.*, 2019] shows that DTW_I has a statistically significant lower performance than MLSTM-FCN and WEASEL+MUSE. Furthermore, DTW_I supports its predictions with limited information (Information: *Features+Time*) that needs to be analyzed by a domain expert to ensure that it is relevant for the application (User: *Domain Expert*). However, DTW_I is an easy-to-understand model (Comprehensibility: *White-Box*) which provides faithful explanations (Faithfulness: *Perfect*) for each MTS (Granularity: *Local*).

Then, we can analyze MLSTM-FCN and WEASEL+MUSE together. First, based on predefined train/test splits and an arithmetic mean of the accuracies, MLSTM-FCN exhibits the best performance on the 35 public datasets

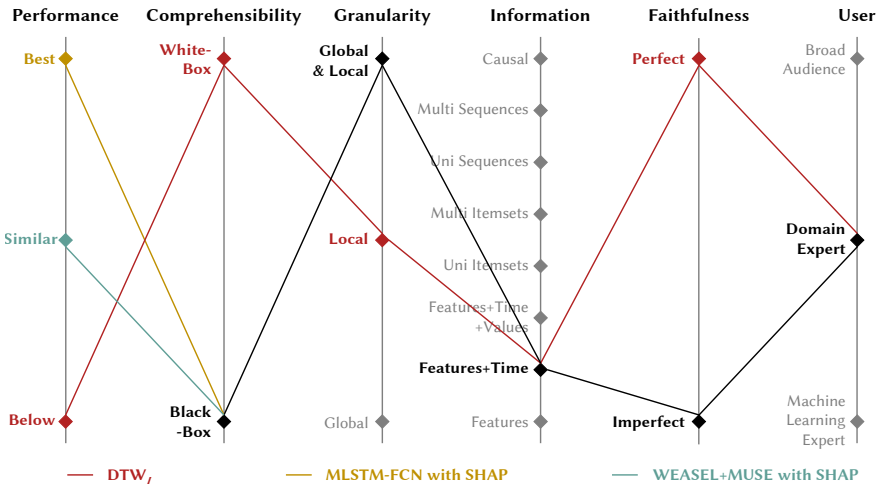

Figure 2: Parallel coordinates plot of the state-of-the-art MTS classifiers. Performance evaluation method: predefined train/test splits and an arithmetic mean of the accuracies on 35 public datasets [Karim et al., 2019]. As presented in section 2.2, the models evaluated in the benchmark are: DTW$_D$, DTW$_I$, gRSF, LPS, MLSTM-FCN, mv-ARF, SMTS, UFS and WEASEL+MUSE.

(Performance: *Best*) [Karim *et al.*, 2019], followed by WEASEL+MUSE (Performance: *Similar*). Second, both MLSTM-FCN and WEASEL+MUSE are "black-box" classifiers without being explainable-by-design or having a post-hoc model-specific explainability method. Thus, the explainability characteristics of these models depend on the choice of the post-hoc model-agnostic explainability method. We have selected SHapley Additive exPlanations (SHAP) [Lundberg and Lee, 2017], a state-of-the-art post-hoc model-agnostic explainability method offering explanations at all granularity levels. SHAP method measures how much each variable (Features+Time) impacts predictions and comes up with a ranking of the variables which could be exploited by domain experts. The combination of MLSTM-FCN and WEASEL+MUSE with SHAP enables them to outperform DTW$_I$ while reaching explanations with a similar level of information (Information: *Features+Time*, DTW$_I$: *Features+Time*), in the meantime remaining accessible to the same user category (User: *Domain Expert*, DTW$_I$: *Domain Expert*). However, as opposed to DTW$_I$, SHAP relies on a surrogate model which cannot provide perfectly faithful explanations (Faithfulness: *Imperfect*, DTW$_I$: *Perfect*).

Therefore, based on the performance-explainability framework introduced, if a "white-box" model and perfect faithfulness are not required, it would be preferable to choose MLSTM-FCN with SHAP instead of the other state-of-the-art MTS classifiers on average on the 35 public datasets. In addition to its better level of performance, MLSTM-FCN with SHAP provides the same level of information and at all granularity levels.

However, the imperfect faithfulness of the explanations could prevent the use of MLSTM-FCN with a surrogate explainable model on numerous applications. In addition, the level of information provided to support the predictions remains limited (Information: *Features+Time*). Therefore, based on the assessment of the current state-of-the-art MTS classifiers with the framework proposed, it would be valuable for instance to design some new high-performing MTS classifiers which provide faithful and more informative explanations. For example, it could be interesting to work in the direction proposed in [Fauvel *et al.*, 2020b]. It presents a new MTS classifier (XEM) which reconciles performance (Performance: *Best*) and faithfulness while providing the time window used to classify the whole MTS (Information: *Uni Sequences*). XEM is based on a new hybrid ensemble method that combines an explicit approach to handle the bias-variance trade-off and an implicit approach to individualize classifier errors on different parts of the training data [Fauvel *et al.*, 2019]. Nevertheless, the explanations provided by XEM are only available per MTS (Granularity: *Local*) and the level of information could be further improved. As suggested by the authors, it could be interesting to analyze the time windows identified for each class to determine if they contain some common multidimensional sequences (Information: *Multi Sequences*, Granularity: *Both Global & Local*). These patterns could also broaden the audience as they would summarize the key information in the discriminative time windows.

5 Conclusion

We have presented a new performance-explainability analytical framework to assess and benchmark the machine learning methods. The framework details a set of characteristics that systematize the performance-explainability assessment of machine learning methods. In addition, it can be employed to identify ways to improve current machine learning methods and to design new ones. Finally, we have illustrated the use of the framework by benchmarking the current state-of-the-art MTS classifiers. With regards to future work, we plan to further elaborate the definition of the different components of the framework (especially the *Model Comprehensibility*, the *Information Type* and the *Faithfulness*) and evaluate the relevance of integrating new components. Then, we plan to apply the framework extensively to assess the different types of existing machine learning methods.

Acknowledgments

This work was supported by the French National Research Agency under the Investments for the Future Program (ANR-16-CONV-0004) and the Inria Project Lab "Hybrid Approaches for Interpretable AI" (HyAIAI).

References

[Altmann *et al.*, 2010] A. Altmann, L. Tolosi, O. Sander, and T. Lengauer. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[Ancona *et al.*, 2018] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross. Towards Better Understanding of Gradient-Based Attribution Methods for Deep Neural Networks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[Bagnall *et al.*, 2018] A. Bagnall, J. Lines, and E. Keogh. The UEA UCR Time Series Classification Archive. 2018.

[Baydogan and Runger, 2014] M. Baydogan and G. Runger. Learning a Symbolic Representation for Multivariate Time Series Classification. *Data Mining and Knowledge Discovery*, 29(2):400–422, 2014.

[Baydogan and Runger, 2016] M. Baydogan and G. Runger. Time Series Representation and Similarity Based on Local Autopatterns. *Data Mining and Knowledge Discovery*, 30(2):476–509, 2016.

[Bohlender and Kóhl, 2019] D. Bohlender and M. Kóhl. Towards a Characterization of Explainable Systems. *ArXiv*, 2019.

[Demšar, 2006] J. Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[Doshi-Velez and Kim, 2017] F. Doshi-Velez and B. Kim. Towards a Rigorous Science of Interpretable Machine Learning. *ArXiv*, 2017.

[Du *et al.*, 2020] M. Du, N. Liu, and X. Hu. Techniques for Interpretable Machine Learning. *Communications of the ACM*, 2020.

[Fauvel *et al.*, 2019] K. Fauvel, V. Masson, É. Fromont, P. Faverdin, and A. Termier. Towards Sustainable Dairy Management - A Machine Learning Enhanced Method for Estrus Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[Fauvel *et al.*, 2020a] K. Fauvel, D. Balouek-Thomert, D. Melgar, P. Silva, A. Simonet, G. Antoniu, A. Costan, V. Masson, M. Parashar, I. Rodero, and A. Termier. A Distributed Multi-Sensor Machine Learning Approach to Earthquake Early Warning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020.

[Fauvel *et al.*, 2020b] K. Fauvel, É. Fromont, V. Masson, P. Faverdin, and A. Termier. XEM: An Explainable-by-Design Ensemble Method for Multivariate Time Series Classification. *ArXiv*, 2020.

[Flach, 2019] P. Flach. Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019.

[Guidotti *et al.*, 2018] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Survey*, 2018.

[Guidotti *et al.*, 2019] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini. Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intelligent Systems*, 34(6):14–23, 2019.

[Hall *et al.*, 2019] M. Hall, D. Harbone, R. Tomsett, V. Galetic, S. Quintana-Amate, A. Nottle, and A. Preece. A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems. In *Proceedings of the IJCAI Workshop on Explainable Artificial Intelligence*, 2019.

[Henin and Métayer, 2019] C. Henin and D. Le Métayer. Towards a Generic Framework for Black-Box Explanation Methods. In *Proceedings of the IJCAI Workshop on Explainable Artificial Intelligence*, 2019.

[Jiang *et al.*, 2019] R Jiang, X. Song, D. Huang, X. Song, T. Xia, Z. Cai, Z. Wang, K. Kim, and R. Shibasaki. DeepUrbanEvent: A System for Predicting Citywide Crowd Dynamics at Big Events. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[Karim *et al.*, 2019] F. Karim, S. Majumdar, H. Darabi, and S. Harford. Multivariate LSTM-FCNs for Time Series Classification. *Neural Networks*, 116:237–245, 2019.

[Karlsson *et al.*, 2016] I. Karlsson, P. Papapetrou, and H. Boström. Generalized Random Shapelet Forests. *Data Mining and Knowledge Discovery*, 30(5):1053–1085, 2016.

[Lakkaraju *et al.*, 2017] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec. Interpretable and Explorable Approximations of Black Box Models. In *Proceedings of the KDD Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[Li *et al.*, 2018] Jia Li, Y. Rong, H. Meng, Z. Lu, T. Kwok, and H. Cheng. TATC: Predicting Alzheimer's Disease with Actigraphy Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

[Lipton, 2016] Z. Lipton. The Mythos of Model Interpretability. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*, 2016.

[Lundberg and Lee, 2017] S. Lundberg and S. Lee. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[Miller, 2019] T. Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267:1–38, 2019.

[Neath and Surprenant, 2003] I. Neath and A. Surprenant. *Human Memory: An Introduction to Research, Data, and Theory*. Thomson/Wadsworth, 2003.

[Palczewska *et al.*, 2013] A. Palczewska, J. Palczewski, R. Robinson, and D. Neagu. Interpreting Random Forest Models Using a Feature Contribution Method. In *Proceedings of the 14th IEEE International Conference on Information Reuse Integration*, 2013.

[Pearl and Mackenzie, 2018] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[Ribeiro *et al.*, 2018] M. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

[Rudin, 2019] C. Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1:206–215, 2019.

[Schäfer and Leser, 2017] P. Schäfer and U. Leser. Multivariate Time Series Classification with WEASEL + MUSE. *ArXiv*, 2017.

[Selvaraju *et al.*, 2019] R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128:336–359, 2019.

[Seto *et al.*, 2015] S. Seto, W. Zhang, and Y. Zhou. Multivariate Time Series Classification Using Dynamic Time Warping Template Selection for Human Activity Recognition. In *Proceedings of the IEEE Symposium Series on Computational Intelligence*, 2015.

[Shokoohi-Yekta *et al.*, 2017] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh. Generalizing DTW to the Multi-Dimensional Case Requires an Adaptive Approach. *Data Mining and Knowledge Discovery*, 31:1–31, 2017.

[Tomsett *et al.*, 2018] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty. Interpretable to Whom? A Role-Based Model for Analyzing Interpretable Machine Learning Systems. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*, 2018.

[Tuncel and Baydogan, 2018] K. Tuncel and M. Baydogan. Autoregressive Forests for Multivariate Time Series Modeling. *Pattern Recognition*, 73:202–215, 2018.

[Ventocilla *et al.*, 2018] E. Ventocilla, T. Helldin, M. Riveiro, J. Bae, and N. Lavesson. Towards a Taxonomy for Interpretable and Interactive Machine Learning. In *Proceedings of the IJCAI Workshop on Explainable Artificial Intelligence*, 2018.

[Wistuba *et al.*, 2015] M. Wistuba, J. Grabocka, and L. Schmidt-Thieme. Ultra-Fast Shapelets for Time Series Classification. *ArXiv*, 2015.

[Witten *et al.*, 2016] I. Witten, E. Frank, M. Hall, and C. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series. 2016.

[Wolf, 2019] C. Wolf. Explainability Scenarios: Towards Scenario-Based XAI Design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019.