Related work

4.1 LIME adaption to time series

LIMESegment: Meaningful, Realistic Time Series Explanations

To adapt LIME to time series classification, Torty and Peter framed three non-trivial challenges without satisfying solutions so far: the meaningful interpretable representation of a time series, proper realistic perturbation to generate new samples, and the definition of a local neighbor for the explained time series sample. The article proposed solutions to these questions by designing Nearest Neighbor Segmentation and Realistic Background Perturbation as well as employing Dynamic Time Warping. These process combined with Linear Ridge Regression model as surrogate model is put together into the LIMESegment. The designed segmentation and perturbation process proves to outperform the existing solutions in finding homogeneous regions and the realism of perturbations, and the LIMESegment has been shown to produce more Faithful and Robust explanations than the existing state-of-the-art adaptation of LIME to TS. The possible space for optimization lies in the improvement of local TS generation and the multivariate adaption following this design.

SEGAL time series classification — Stable explanations using a generative model and an adaptive weighting method for LIME

By investigating the stability of LIME when adapted to explaining multivariate TSC problems, Han et al. put emphasis on research highlighting challenges around the explanations. In some cases, the explanations provided vary over repeated runs of the algorithm, which proves to be the influence of the out-of-distribution problem induced by traditional neighbour generation methods. The solution proposed is a two-fold approach, improving both the traditional neighboring sample generation process and the weighting strategy. A generative model is designed to prevent out-of-distribution problems, and a newly designed Adaptive weighting method which adaptively allocates weights to closer neighbors and is less considerate of l the absolute distance. These approaches proved to have made LIME explanations more stable across repititions through datasets. The sole challenge raised by the article is that the feature importance of time series may not be understandable enough in terms of the general utility of explanations, and other means of explanations is needed in the future, such as counterfactual explanations.

**B-LIME: An Improvement of LIME for Interpretable Deep Learning Classification of Cardiac Arrhythmia from ECG Signals**

The article propose a B-LIME technique that improves LIME to explain signal data, taking into account the special temporal dependency between features that matters more for time series data compared with tabule and images. As a domain-specific XAI technique, B-LIME have made improvements to existing LIME adaptions in three steps throughout the process, including neighbor generation technique, explanation method, and means to demonstrate the explanation to make the explanations meaningful, credible and understandable. The B-LIME explanations were examined in a proposed hybrid CNN-GRU model for cardiac arrhythmia prediction and compared with LIME explanations. In comparison to LIME, which highlights random areas, B-LIME performs well in highlighting key areas that physicians normally use to diagnose cardiac arrhythmias, such as the QRS complex. The possible future extension includes a deeper and more refined investigation of the model’s behavior, such as the reason the model classifies a prediction as class A but not class B, and the attempt to apply B-LIME to a more signal-based model and evaluate its performance compared to LIME.

**TS-MULE: Local Interpretable Model-Agnostic Explanations for Time Series Forecast Models**

As relatively early research about time series adaption of LIME, the article mainly focused on the segmentation of time series. The existing fixed-width window segmentation may fail to capture the global relationship inside time series, thus making meaningful segmentation a challenge. TS-MULE extends LIME by introducing five novel segmentation algorithms tailored for time series, incorporating techniques like matrix profiles, SAX transformations, and adaptive windowing. These methods improve local attributions and address challenges with fixed segmentation strategies. The evaluation metric is based on CNNs, RNNs, and DNNs, which demonstrated better fidelity in perturbation analysis compared to random baselines. This early adaption is meaningful while still have limitations include dependency on dataset characteristics and model performance, suggesting future comparisons with advanced XAI techniques like SHAP and the integration of domain-specific shapelet discovery.

**A Comprehensive Explanation Framework for Biomedical Time Series Classification**

This study addresses the challenge of interpreting deep learning (DL) models for biomedical time series classification, focusing on atrial fibrillation (AF) detection from single-lead electrocardiograms (ECGs). The authors propose a post-hoc framework combining global and local explanation techniques. Globally, they analyze class-level data to identify input features, such as R-R interval variability and P-wave absence, crucial for model decisions. Locally, they highlight signal regions triggering specific outputs. Using ablation studies, permutation analyses, and LIME, they assess the interpretability of MobileNet, demonstrating its alignment with clinical expertise. The framework identifies physiologically meaningful features while providing explainable insights even in misclassified cases, often tied to real signal anomalies. Testing on the PhysioNet dataset shows robust accuracy (84.38%) and reliability in explaining predictions. Future directions include validating this framework in clinical trials and extending it to other biomedical applications, fostering trust in DL use for sensitive health diagnostics.

Interpretable heartbeat classification using local model-agnostic explanations on ECGs

To address the interpretability challenge in ECG time series classification, Neves et al. developed a novel framework that adapts model-agnostic explainable artificial intelligence (XAI) methods. Key innovations include introducing temporal dependency into explanations through the integration of amplitude and derivative signals, and adapting methods like Permutation Sample Importance (PSI) and Local Interpretable Model-Agnostic Explanations (LIME) for ECG data. The framework also employs validation techniques such as Jaccard Index and performance decrease analysis to measure the faithfulness of explanations.

The proposed approach demonstrates superior performance compared to standard XAI methods like SHAP, particularly in temporal sensitivity and robustness. A user study highlights its utility in aiding clinical decision-making and training. While promising, the study acknowledges future improvements in model scalability and multivariate time series adaptation. This work contributes a significant step towards trustworthy and interpretable ML applications in cardiovascular diagnostics.

VAE-LIME: Deep Generative Model Based Approach for Local Data-Driven Model Interpretability Applied to the Ironmaking

Local interpretable model-agnostic explanations (LIME) are widely used for interpreting black-box models, yet they face limitations in fidelity and stability. In response, VAE-LIME leverages a Variational Autoencoder (VAE) to enhance local interpretability for multivariate time series data. Unlike traditional LIME, which relies on random perturbations, VAE-LIME generates realistic samples by learning latent space distributions from the training data. This approach improves the local fidelity of surrogate models and the stability of variable importance metrics. Applied to blast furnace temperature prediction, VAE-LIME significantly reduces mean squared errors (MSE) and improves R² scores compared to LIME. These enhancements are attributed to the controlled sample generation that respects variable correlations inherent in the process. While results highlight the effectiveness of VAE-LIME, further research is needed to optimize temporal stability and extend applications to inherently interpretable models. The methodology offers a promising direction for interpretable AI in industrial contexts.

4.2 Perturbation analysis: what constitudes of a good perturbation

As mentioned above, adapting LIME on TSC problems has been a relatively new domain, therefore research specifically focusing on optimizing the perturbation procedure in neighbor generating is not abundant. However, perturbation methods for other purposes but with similar principles have been studied in other areas.

A Deep Dive into Perturbations as Evaluation Technique for Time Series XAI

Methodology:

The aim of the improvement is to properly identify the common background component behind all samples. To solve the possible class-specific pattern dominance problem, we draw inspiration from conventional statistical methods. As a basic way to seek to achieve a fair and unbiased representation of a group, the first considered sampling method is simple random sampling, which in terms of TS background identification method, is to concatenate the time-frequency representation of every shifted randomly selected sample (or if the dataset is big enough, all samples), and find the most persistent frequency from the matrix. Detailed process is represented as follows:

Another sampling approach which can be referred to is the stratified sampling. In the original background perturbation approach, the TS samples are discreted into various frequencies, making it a composition of harmonic oscillations, therefore the problem can be seen as choosing the most representative oscillation inside a sample among all samples. By treating samples as subgroups of the whole, each sample can be applied a background extraction procedure and generate a ‘candidate’ background, and the combination of these ‘candidates‘ can be seen as a time-frequency representation that contains underlying information from all samples. The candidates group is then applied a background identification procedure again to generate the final background. The idea of treating each sample as subgroups came from stratified sampling, but unlike the sampling procedure, the oscillations were chosen by its persistency and variance instead of randomly to meet the need for backgrounds.

Experiment:

Based on the theoretical analysis and the initial idea of the design, which is to prevent class-based features surpassing the background feature in realistic time series data, the evaluation of the MRBP is tested mainly on multiple realistic time series data. Although synthetic datasets were used in the LIMESegment evaluation, classwise difference is contained solely in the last 20% of the time series, and therefore the above problem that may occur in real data is avoided. In order to prevent the bias created by synthetic datasets and to ensure various realistic situations are considered, the experiments will be carried out on multiple time series datasets of all types.

4.1 evaluating the quality of perturbation

To evaluate the quality of background contents generated by MRBP, the experiments designed to evaluate RBP is followed, originating from the intuition of Agarwal and Nguyen (2020) adapted to TS: after applying a perturbation to a test set, the more successful the perturbation, the worse the classification performance. The Considering the pre-constructed synthetic dataset problems, the test sets used is changed to realistic datasets. We select a 1D Convolutional Neural Network as our black box, and apply the normal time series classification procedure to gain the original model together with its performance. For each sample in the original test set, a time segment whose length is proportional to the overall time series will be perturbated with either RBP, MRBP, blurring, zeroed, or random values. These perturbated datasets will be obtained classification accuracy, and a more significant accuracy decrease indicates a more successful perturbation and background content of a higher quality.

Due to the uncertainty for ground truth of real TS datasets, the experiment is implemented on 10 datasets of various domains including Audio, Motion, Human Activity Recognition, etc., and the hypothesis and purpose in the design held true for most of the results. The experiments also control the size of datasets, proportion of the perturbated segment, length of time series and number of classes, some of which influenced the accuracy decrease in a meaningful way.

The first feature to notice is that the overall accuracy, including the original and perturbed ones, were not as ideal as the previous experiments using synthetic datasets. The accuracy varies through datasets, most of which were not state-of-the-art for a TSC model, unlike the experiment for RBP, whose original accuracy was 1.0. Such phenomenon proved the complexity of the real data set, as the basic CNN model were usually used as a baseline model for TSC research. RBP did not perform as good as it did on synthetic dataset in the previous literature. The results show that the accuracy after RBP was not always significantly lower than the blurring, zeroed, or random values. Such performance may indicate that for some samples, the background content used by RBP may not be the most uninformative one.

For a large proportion of the datasets, the MRBP perturbation shows a more significant decrease in accuracy compared with RBP, zeroed and blurring, proving a generally better background content from MRBP. The Noise perturbation showed a lower accuracy in some datasets, however the following experiments proved that the noise perturbation is seriously flawed on certain aspect and cannot be considered as a realistic perturbation.

As mentioned before, during the experiment, the datasets used were diverse in length of time series and number of classes, and to further ensure faithfulness of the experiment, proportions of the perturbated part were adjusted to observe its influence. The results shows that among these parameters, the number of classes and the proportions of the perturbated part has a regular influence on results of some or all perturbation approaches, which may reveal features of these perturbations.

In general, when the proportions of the perturbated part is low, the accuracy decrease from each perturbated datasets were less significant, even negligible in some datasets when the proportion is small enough, and the superiority of each method tend to be less stable. This result is intuitive: the more the perturbated features, the less classwise information the example includes. When facing small perturbations, the model shows a slight anti-adversary capability. Such result assures that perturbations have been applied to segments that indeed contains information.

The number of classes also influences test results, but not in the same degree for all sorts of perturbations. Generally, due to the limited performance of the model, it tends to perform worse on datasets with more classes, but the gap in the evaluation metrics of RBP and MRBP seems to be wider for datasets with more classes. This may indicate a comparatively better performance for MRBP, as the gap between performances of perturbations should be smaller as the original accuracy decreases. After analyzing the statistical basis of both methods, a possible explanation for this is that for multiclass datasets, time series of each class accounts for a smaller proportion of all samples, therefore the background content will be concatenated several times more than each classwise content, making it more likely to be selected undisturbed.

To sum up, the perturbation quality test results show the difference between generating background contents for synthetic datasets and real ones, and confirms that MRBP did make an improvement based on RBP to adapt with classwise pattern dominance problems in real time data.

4.2 Realism Test

4.2.1 Model Separation Test

To evaluate whether the TS perturbed RBP produces more realistic TS, the experiment from LIMESegment article is followed, whose theory is based on adversarial attack evaluations from Chen et al. (2020). A 1D CNN is used as test model trained on a binary class synthetic TS dataset, which contains unperturbed and perturbated data, labeled 1 and 0. During the training process, the validation loss curve for varying perturbation strategies including RBP, MRBP, blurring, zeroed, random will show how quickly and successfully the model learns and generalizes differences between these two classes.

The results are displays in Figure 5, which was not ideal. The MRBP approach did not successfully confuse the model like RBP did, which may indicate that MRBP is less realistic. The reason for the poor performance might be that the background contents MRBP generates are the same across samples, which is relatively easier for ML models to recognize compared with sample-specific perturbation methods. This may not be intrusively right, as an ideal background content should be universal across all samples. To further evaluate the realism of MRBP, other approaches evaluating the realism and reasonability of perturbations were researched and applied.

4.2.2 L2-Norm Test

Most of the perturbation analysis researches do not put much emphasis on the realism of perturbations , therefore another realism evaluation metric in Gautier et al.(2020), originating from the L∞ norm commonly used in the case of attacks on images, but when adapted to time series, the L2 norm is required due to the importance of global properties of time series. The mean L2 norm