Improving LIMESegment through Multisample adaption of the Realistic Background Perturbation

Introduction

The application of machine learning models has significantly transformed various fields, from healthcare to finance, enabling automation and enhancing decision-making processes. Despite these advancements, the adoption of machine learning, especially complex models like deep neural networks, has encountered several challenges, primarily related to transparency and explainability. These issues are crucial as many machine learning models, often regarded as "black boxes," make decisions that are difficult for users to interpret or understand. Such opacity raises concerns regarding trust, fairness, and ethical use, particularly in critical domains like healthcare and law enforcement. Therefore, the necessity for transparency has led to the development of eXplainable Artificial Intelligence (XAI) techniques, which aim to bridge the gap between complex machine learning behavior and human understanding.

XAI is motivated by the need to make AI systems more interpretable, ensuring that their decision-making processes can be understood, trusted, and properly audited by humans. The concept of XAI is grounded in various techniques designed to provide insights into the inner workings of machine learning models. Its motivations stem from the demand for fairness, accountability, and the need to comply with regulations that require AI systems to be explainable. XAI techniques are increasingly applied across multiple domains, such as healthcare, finance, and autonomous driving, to provide stakeholders with understandable explanations for model predictions, thereby building trust and facilitating informed decision-making.

One prominent XAI method is Local Interpretable Model-agnostic Explanations (LIME). LIME focuses on generating locally interpretable explanations for the predictions made by complex models, regardless of the model type or architecture. It does so by approximating the model with a simpler, interpretable model in the vicinity of the instance being explained. This local model-agnostic approach makes it possible to derive explanations without requiring knowledge of the inner workings of the model, thus offering flexibility and ease of use. In this context, understanding the structure of arguments and providing explanations for the relationships between components are particularly important for establishing a coherent and understandable justification for predictions.

**Argument Structure and Local Model-Agnostic Explanations**

In argument mining, understanding and mapping the relationships between claims, evidence, and supporting arguments are essential tasks that help construct logical and transparent arguments. Argument mining involves several subtasks, such as component segmentation, component classification, and link prediction. Each of these tasks aims to identify, classify, and establish relations between different parts of an argument, ultimately creating a structured representation of the argument. This structured representation is not only crucial for constructing logical arguments but also serves as the foundation for generating explanations for machine learning predictions.

Local model-agnostic explanations, such as those provided by LIME, focus on approximating the behavior of a machine learning model within a small neighborhood around the data instance being explained. This approach aims to find a simple, interpretable representation that can explain why a particular instance received a specific prediction. LIME does this by perturbing the input data and observing the model's response, thereby creating a local linear model that can explain the instance in question. This method has proven to be effective in providing explanations for models used in natural language processing, image recognition, and time series analysis.

Related work

**Locally Interpretable Model-Agnostic Explanations**

The background of this article starts with the Locally Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016), which is the basic form of our XAI technique.

LIME’s goal is to identify an interpretable model over the interpretable representation that is locally faithful to the classifier. Given as input a dataset X, the black box classifier f : Rd → R, the instance to be explained x ∈ Rd its associated classification Yx ∈ R, LIME applies a mapping from the original instance x into its interpretable representation σ(x). If x is an image, σ(.) is often a grouping over the original pixel space into human interpretable “super pixels”. σ(x) = 1 d ′ can therefore be understood as a vector of ones to indicate each super pixel is “turned on” in the instance to be explained where d ′ corresponds to the number of “super pixels” (d ′ < d). LIME then generates new samples σ(z) by randomly “turning off” dimensions of the interpretable representation through drawing nonzero elements of σ(x) uniformly at random. LIME converts the set of generated samples σ(Z) into the original feature space to obtain the labels YZ = f(Z). It approximates the behavior of f with an explanation model, g : Rd ′ → R. Commonly, a linear model such as ridge regression is selected as g. An exponential kernel Kexp(σ(x), σ(Z)) is used to weight g such that generated samples which are more similar to the instance to be explained have more of an effect on the resulting explanation. LIME uses the coefficients of g, w ∈ Rd ′ to be used as an explanation of x.

Transferring LIME into Time Series Classification Explanation

Attempts to use LIME to generate both univariate and multivariate time series classification problem explanations has risen in the former five years, and is developing from primitive adaptions to improved ones. 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. (**TS-MULE: Local Interpretable Model-Agnostic Explanations for Time Series Forecast Models**) Another attempt comes from Neves et al. developing a novel framework that adapts model-agnostic explainable artificial intelligence (XAI) methods to address the interpretability challenge in ECG time series classification. 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. (**Interpretable heartbeat classification using local model-agnostic explanations on ECGs**)

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**

**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

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.

Model agnostic interpretability of machine learning.

4.2 Perturbation analysis: what constitutes 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:

This chapter will cover the approaches used to improve the XAI.

The first section will begin with the basic technique which our research is based on, the LIMESegment, in which Our research emphasis on the Real Background Perturbation process of the LIMESegment. This will lead to the reveal of possible bias and instability from the perturbation process.

The next section contains novel perturbation approaches aiming to reduce the instability problem of the original perturbation method. Two optimization schemes were proposed and justified their pros and cons.

The last section focuses on identifying possible flaws of the design when facing atypical datasets, and corresponding optimizations.

3.1 LIMESegment

LIMESegment is an adaptation of LIME (Locally Interpretable Model-Agnostic Explanations) designed specifically for time series classification (TSC). The framework addresses three key challenges: meaningful segmentation, realistic perturbation, and robust similarity measurement. Its three main components—NNSegment, RBP, and Dynamic Time Warping (DTW)—are combined to generate interpretable and reliable explanations for time series models.

The first step in LIMESegment involves transforming the time series into interpretable segments called "super segments." Unlike traditional methods that use arbitrary fixed-length windows, NNSegment identifies segments based on changes in shape and statistical properties: The time series is divided into overlapping windows. Each window is compared to its neighbors using normalized cross-correlation to identify similarity. Breakpoints are detected when adjacent windows show differences in mean and variance, indicating a behavioral change in the series. These super segments are considered features of the time series sample.

The second step generates perturbations for the segmented time series to simulate a local neighborhood. LIMESegment introduces the Realistic Background Perturbation (RBP) method to ensure that the perturbations are both natural and realistic: Using the Short-Time Fourier Transform (STFT), RBP identifies the most persistent and stable frequency components in the series. These components represent the "background" signal.

After generating neighbors of the explained sample, LIMESegment employs Dynamic Time Warping (DTW) to measure the similarity between the original time series and perturbed samples. It is used to calculate distances between the original instance and perturbed samples. These distances are normalized and transformed into weights through an exponential kernel, which are then applied to the surrogate model. DTW ensures that the most similar samples have the greatest influence on the explanations, resulting in more stable and intuitive outputs.

Among these steps, the Real Background Perturbation approach is where the improvement design lies. It is based on findings from harmonic analysis: Any TS x can be represented as a composition of harmonic oscillations in the frequency domain such that x = xω = R ∞ −∞ xte −2πtωdt. Tω can be viewed as a distribution over the frequency content of the signal. Maxima in the frequency domain reflect a high proportion of the signal oscillating at that frequency. The identification of backgrounds is based on such finding: realistic background content represents a global property of an image and is not necessarily the local low frequency content but the most commonly occurring global frequency information (Agarwal and Nguyen, 2020). To understand how the spectral density of a TS changes over time we use the Discrete Short Time Frequency Transform (STFT). Given a TS x = [x1; ...; xT ] where x is considered a discrete time representation of the underlying phenomenon, the STFT converts x into its time-frequency representation by taking the Fourier transform of x multiplied by a sliding window which is non-zero only for a fixed small length ws: ST F T(x, ws, ω) = Σ∞ −∞xw(T − ws)DF TT where DF TT = e −iωT represents the Fourier transform, ω is 0 200 400 600 800 1000 time −2 −1 0 1 2 amplitude a) original signal 200 400 600 800 time 0.0 0.1 0.2 0.3 0.4 0.5 frequency b) STFT of original signal Figure 4: Intuition behind RBP. Original signal (a) composed of background signal and varying frequency sine waves at indexes: [0 : 100], [400 : 500] and [600 : 800]. b) shows the spectrogram obtained by applying STFT to the original signal. The spectrogram captures the background signal which remains constant through time as well as the shorter length “content” sine waves at their respective frequencies. the frequency parameter and w(.) is a window function parametrized by window size. The STFT, when applied to a discrete TS results in a matrix which records magnitude and phase for each point in time and frequency. We are interested in filtering the signal by selecting only the background content. Given ST F T(x) we therefore find the most persistent frequency by first selecting only the magnitude response as |ft| and then finding the frequency band which has the highest value over time with minimal variance: Fpersist = argmaxf µ(|ft|) σ(|ft|) : f, t ∈ {STFT(x)} where µ(|ft|) σ(|ft|) is the mean magnitude response normalised by its standard deviation of a selected frequency band over time.

Although the RBP approach has its justification, the theory behind did not necessarily ensure the background identified being the intuitive kind of background, which is the universal and the lowest level pattern across sample of the same kind. Instead, the background is the most basic pattern for the interpreted sample, which could not guarantee its universality because no other samples are considered. Such limitation could cause the following situation happening: the class wise content may be significant enough to surpass the background content, causing the approach to identify components with class-based information as the background of the sample. This may lead to perturbations unable to ‘delete’ information as expected when replacing super segments with identified contents, which would eventually lead to underestimation of the importance of some features, as the neighbors would remain more class wise information than it should have. Although the problem may occur only on a small proportion of samples, it still reflects the instability of this method.

3.2 Global Background Identification via Multisample Realistic Background Perturbation

To solve the possible class-specific pattern dominance problem, when applying background identification perturbation, other samples in the dataset should be considered. To seek for a representative background content among all samples, we draw inspiration from conventional statistical methods.

The aim of the improvement is to properly identify the common background component behind all samples. 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:

After the STFT converts x into its time-frequency representation through the procedures above, every Time-Frequency matrix ST F T(x, ws, ω) = Σ∞ −∞xw(T − ws)DF TT from a sample will be added to a global matrix ST F T({x1,x2,……,xn}, ws, ω) = Σn −1Σ∞ −∞xw(T − ws)DF TT. This is made possible due to the same fixed window size and frequency range across all samples, and the need for phase(which cannot be added)magnitude only for this problem. Given the global time-frequency matrix, the Fpersist = argmaxf µ(|ft|) σ(|ft|) is therefore used to find the most persistent and significant frequency. After scaling the Fpersist by dividing with the number of samples used, it is converted into the original time domain via the inverse STFT, from which, the relevant segments of background content can be chosen to replace parts of the original signal.

Another sampling approach which can be referred to is the stratified sampling. In the original background perturbation approach, the TS samples are discrete 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 candidate group is then applied a background identification procedure again to generate the final background.

The approach starts with collecting background contents of all samples in the dataset to gain { Fpersist1, Fpersist2,……, Fpersistn}, in which Fpersistn = argmaxf µ(|ftn|) σ(|ftn|). These background contents will be added to a global matrix Mbackground = Σn 1 Fpersistn ,and is applied background identification again. The Fpersist = argmaxf µ(|ft|) σ(|ft|) : f, t ∈ {STFT(x)} is identified as the most persistent frequency. The conversion into the original time domain is the same with the original approach.

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.

3.3 Stability enhancement for Multisample Realistic Background Perturbation

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.

ECG200

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) is used, originating from the L∞ norm commonly used in the case of attacks on images. When adapted to time series, the L2 norm is required due to the importance of global properties of time series. The mean L2 norm

\section{Comparison Analysis}

According to the research performances above, RBP approach has shown its limitation upon dealing with complicated real datasets in which dominance of backgrounds towards class wise features is not guaranteed. For most datasets and hyperparameters, the MRBP outperforms RBP in terms of the quality of perturbations, and such advantage tends to be more significant on datasets with a larger variety of labels. However, the MRBP showed a weaker performance than RBP both in confusing the CNN model and in L2-norm metric, indicating that MRBP did not successfully maintain realism attribute from its origin approach.

\section{Limitations and Insights}

Although MRBP reached its main purpose of optimization, there are still limitations according to some performance in the tests.

The main problem unsolved of the MRBP is the realism of the generated perturbation, which is inferior to the previous method. According to Torty et al.(2022)[], this may cause salient visible or invisible information contained in the super segment. A possible way of dealing with the problem is to keep the method sample specific as well as consider other samples of the same kind. Although having a common background across samples is intuitive, a sample specific background which considers contexts would make the perturbation smoother, thus making it more realistic.

In addition, Experiments on small datasets tend to perform unstable and less satisfying. This problem arises in all approaches, but solving this would be of great practical interest. Possible solutions including changing the classification model with Meta-Learning Models, or to develop domain-specific techniques with the help of some external conclusions.

\section{Overall Result}

The MRBP approach is proved to have the best performance in terms of the quality of perturbations and is stable across datasets. Although it did not outperform RBP in realism tests, but such attribute were much less considered in similar researches, and MRBP does not lag RBP by much in the l2 norm metric, hence considering both the quality and authenticity of the disturbance, the MRBP can be considered an improvement based on RBP to better adapt to real life data.

Conclusion

5.1 Summary

This section will recap the previous sections and outline the project’s main contributions.

Chapter 1 demonstrates current dilemmas in applying machine learning models and introduces the topic of XAI and its motivations and applications. It explains the argument structure, and the idea of local model-agnostic explanation. A review of the former research and application of the XAI technique is provided, and the most significant challenges of the adaption are listed. Finally, the main contributions of the project and its purpose is outlined, specifically focusing on the justification for the design of the proposed method.

Chapter 2 summarizes related literatures with background research necessary for readers to understand the thesis. It covers the LIME technique and its adaption to time series classification models. Specifically, it covers the general focus of previous research and the approaches adopted to address it. In addition, the chapter also covers papers of different fields related to perturbations to help justify our approaches in the rest of the thesis.

Chapter 3 covers the main design of the project, outlining the design procedure of the MRBP and the realization of the stability improvement based on the original perturbation approach of the LIMESegment. The chapter begins with an overview of LIMESegment, one of the existing time series adaption of LIME that considers all major challenges without using ML models in its process. Specifically, it introduced the theoretical basis and procedure of RBP, and demonstrated the possible flaw inside the methodology which may cause instability. Two designs of improvement on RBP were then introduced, together with fine tuning to solve problems like possible atypical datasets.

Chapter 4 evaluated the proposed method in both effectivity and realism. It started by explaining the effectivity test design as well as analyzing the potential bias of the synthetic dataset used in former research. The e effectivity test is then adapted to real datasets, aiming to prove evidence for how successfully the perturbations cover the information on certain aspects. Another factor considered is the realism of perturbations generated, which contains Model Separation Test and L2-Norm Test. These experiments show that MRBP performs better than RBP in the quality of perturbation, but it lost part of its realism compared to RBP. This may due to sample-agnostic attribute of the MRBP, but was considered less important than effectivity of the perturbation, therefore MRBP is still considered as an improvement

5.2 Project Status

The aim of the project are as follows:

To first understand the theory of LIMESegment to find out potential gaps for improvement.

To extend the work of LIMESegment by improving Realistic Background Perturbation to enable multisample background generation.

To consider the feasibility of generative models being used in neighboring samples being used based on LIMESegment.

To evaluate how the RBP and MRBP performs on the real datasets.

From the points just mentioned, the goals that were not complete will be discussed.

Firstly, although the generative model was used in former research in neighbor generation of LIME, the idea of using opaque machine learning model during the procedure of explaining opaque machine learning model is logically flawed to some extent, especially for generative models which are even harder to interpret compared with TSC models. Even if the generative model is not directly involved in the decision interpretation of the classifier, which would reduce the transparency of explanations, and increases the complexity of the explanation system, making it harder for non-technical people to understand.

For the multivariate adaption part, the plan was initially established based on the significance of what this adaption can achieve, while the feasibility and theory basis was not ensured. There were not many researches about univariate LIME techniques adapting to multivariate, most of which considered multivariate adaptions as their research goals in the first place. As the initial theory basis, a multivariate TS LIME is used as a reference method to achieve this goal. However, after initial trials, this approach clearly falls short of the quality of most current researches in the quality of explanations, because the surrogate model its using loses too much information in the time series. There may be other approaches available for this project goal, but due to limited project time, it was not introduced in the paper and will be covered in the future work instead.

Further Work

Further improvement in the MRBP approach

As mentioned in chapter 4, although MRBP is an improvement in general, it loses part of the realism compared with the original design, therefore more work can be done to make the perturbation approach more realistic. According to related research like ML-based neighboring generation, most perturbations are sample-specific, which means that different perturbations were used for different explained samples. In this case, the generative model considers both global patterns of the dataset and the context of the segment perturbated. For MRBP, it is possible to achieve a similar effect by changing statistical methods and applying data processing techniques.

A relatively simple way to add sample-specific parameter is by assigning weights to different samples during background identification process. In practice, a relatively higher weight can be assigned to the specific sample, and determining the specific weight requires a certain statistical theory as a support or obtained through repeated experiments. Such method should contribute to a smoother perturbation, and if the result still requires additional smoothing technique, insights can be drawn from similar domains like adversarial attacks.

Multivariate time series extension for LIMESegment

Multivariate time series classification models provide a better solution for computational tasks on more detailed data like biomedical and industrial time series data. Understanding these models is not only a major problem in the field of TSC, but also a much more complex topic, as it requires capturing the interrelationships between different variables.

Although insightful research about LIME extending to multivariate time series is hard to find, for other XAI techniques, there were some transferrable methods proposed. In (Evaluating Explanation Methods for Multivariate Time Series Classification), SHAP is adapted to multivariate TSC models by modifying multivariate datasets. One of the methods is to connect multiple time series representing different features into a single time series, and the other one is to train and explain one model for each channel independently. The MTSC model in this case is an ensemble of per-channel UTSC models. Theoretically, Both designs can be applied to LIMESegment, but multiple evaluations are required to observe whether the order of the input features or the connection between two feature series would influence the explanation results.