Unsupervised Multi-sensor Anomaly Localization with Explainable AI

Mina Ameli[1,2](✉), Viktor Pfanschilling[3], Anar Amirli[1,2], Wolfgang Maaß[1,2], and Kristian Kersting[3]

[1] Saarland University, Saarbrücken, Germany
[2] German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
{mina.ameli,anar.amirli,wolfgang.maass}@dfki.de
[3] TU Darmstadt, Darmstadt, Germany
{viktor.pfanschilling,kersting}@cs.tu-darmstadt.de

Abstract. Multivariate and Multi-sensor data acquisition for the purpose of device monitoring had a significant impact on recent research in Anomaly Detection. Despite the wide range of anomaly detection approaches, localization of detected anomalies in multivariate and Multi-sensor time-series data remains a challenge. Interpretation and anomaly attribution is critical and could improve the analysis and decision-making for many applications. With anomaly attribution, explanations can be leveraged to understand, on a per-anomaly basis, which sensors cause the root of anomaly and which features are the most important in causing an anomaly. To this end, we propose using saliency-based Explainable-AI approaches to localize the essential sensors responsible for anomalies in an unsupervised manner. While most Explainable AI methods are considered as interpreters of AI models, we show for the first time that Saliency Explainable AI can be utilized in Multi-sensor Anomaly localization applications. Our approach is demonstrated for localizing the detected anomalies in an unsupervised multi-sensor setup, and the experiments show promising results. We evaluate and compare different classes of saliency explainable AI approach on the Server Machine Data (SMD) Dataset and compared the results with the state-of-the-art OmniAnomaly Localization approach. The results of our empirical analysis demonstrate a promising performance.

Keywords: Anomaly localization · Explainable artificial intelligence · Unsupervised anomaly detection · Multivariate time-series · Multi-sensor data

1 Introduction

Anomaly detection has been an active topic in diverse research communities. There is a wide range of existing approaches to this end. With the increasing amount of Multivariate data generated in many monitoring applications and domains such as process industry, medical diagnosis, computer networks,

Published by Springer Nature Switzerland AG 2022
I. Maglogiannis et al. (Eds.): AIAI 2022, IFIP AICT 646, pp. 507–519, 2022.
https://doi.org/10.1007/978-3-031-08333-4_41

etc., unsupervised anomaly detection on multivariate data is of great importance. While many solutions focus on the accuracy of the anomaly detection approaches, localizing the root cause of anomaly for decision-making, analysis, and interpretability of the proposed solutions are critical. To this end, anomaly detection can be considered as a pre-processing step for anomaly localization. Anomaly detection in multi-sensor setup refers to identification of rare events in specific time and anomaly localization refers to identification of the sensors which caused a particular anomaly. There have been several studies that, in particular, focused on the interpretability of predictions in multivariate time-series, including [6,14,23]. For instance, [15] presents the series saliency framework for temporal interpretation for multivariate time-series forecasting. Some other studies focus on the interpretability and trust-worthiness of classifiers [19], or other AI models like Autoencoders [17]. However, there is presently a lack of approaches for multi-sensor and multivariate time-series data anomaly localization. Currently, most approaches in this scheme focus on anomaly detection, and the few that handle in the wild scenarios do not focus on anomaly localization. Localizing the anomalous sensors in a wide number of sensors can assist reliable decision making in monitoring systems. Furthermore, with the explainability of AI solutions attracting more the attention of researchers and being used for different purposes, there is the opportunity to not only interpret the models with this tool but to use it as an additional application in solution pipelines. This paper, proposes using Saliency Explainable AI approaches to localize anomalies in multivariate time-series, and our main contribution can be summarized as follows:

- A novel method for unsupervised Multi-sensor Anomaly Localization.
- A novel application of saliency Explainable AI methods.
- Literature review over available multivariate anomaly localization solutions.
- Evaluation the performance of different Explainable AI methods for anomaly location on multi-sensor and multivariate data.
- A generalized Anomaly Localization solution for different kinds of Anomaly Detection models.
- Proposing future focus of studying on the impact of explanations on bridging the gap between AI practitioners and domain experts.

The remainder of the paper is organized as follows. In Sect. 2, we outline the related work. The Background is explained in Sect. 3. The proposed solution is addressed in Sect. 4. Experimental results are presented in Sect. 5. In Sect. 6, the results are discussed. And Sect. 7 concludes the paper.

2 Related Work

Usage of Anomaly Detection and solutions to address this subject in different multivariate data applications have a long history such as TadGAN [7], LSTM [8], DAGMM [28], and OmniAnomaly [20], but the focus of literature review in this paper is anomaly localization. In this section, we introduce different kinds of anomaly localization that already were investigated in multivariate data in the relevant literature. Feature importance, feature attributions, root cause analysis, and some other terminologies could be found in the literature to address the anomaly localization of multi-sensor, multivariate datasets, and data streams:

[9] proposes sparse PCA methods to perform anomaly detection and localization for network data streams. They identify a sparse low dimensional space that captures the abnormal events in data streams to localize anomalies.

[3] proposed a feature importance evaluation approach which is designed for Isolation Forests. [13] performs this task by examining the contribution of each dimension individually to the decision statistics.

[18] uses ARCANA, an autoencoder-based anomaly root cause analysis. It describes the process of reconstruction as an optimization problem that aims to remove anomalous properties from an anomalous instance.

The anomaly interpretation solution from [20] is to annotate the detected entity rare-event with the top few univariate time-series ranked by their reconstruction probabilities. Recently, [1] extends Shapley Additive Explanation (Shap) to explain anomalies detected by an autoencoder.

[24] proposed a novel anomaly attribution approach for multivariate temporal and spatio-temporal data based on the counterfactual replacements of variables within the anomalous intervals. This counterfactual replacement determines whether the anomaly still has occurred if a subset of variables is more similar to the data outside of that interval.

[4] transform multivariate time-series into the 2D images and with point-wise convolution in a series of images ensures to encode temporal information of each time-series data as well as the correlation between each variable. As a result, an anomaly can be detected and localized by conducting a residual image and an anomaly score function.

Our Anomaly localization method differs from any of the mentioned approaches. While most state-of-the-art anomaly localization approaches are either depending on the anomaly detection pipeline or describe the probability of features contributions by statistical study, we show for the first time that Saliency Explainable AI can be utilized in multi-sensor time-series Anomaly localization applications.

3 Background

In this section, we describe the details of our proposed Multi-Sensor Anomaly Localization with Explainable AI in an unsupervised setup. The overall architecture is shown in Fig. 2. In the following sections, we briefly introduce all the associated fields.

3.1 LSTM Autoencoder

Reconstructions based on LSTM Autoencoders are commonly utilized for unsupervised multivariate anomaly detection and have shown to be effective [11].

As a result, in the anomaly detection pipeline, we used this family of algorithms as the fundamental training architecture. Autoencoders are technically unsupervised learning methods, as they do not require labelled data. For the purpose of anomaly detection, a part of the data which is assumed as normal data and doesn't contain anomalous samples is considered for training, therefore, the method is a semi-supervised method in the context of anomaly detection. An LSTM Autoencoder uses an Encoder-Decoder LSTM architecture to construct an autoencoder for sequential data. In the architecture shown in Fig. 1, a LSTM Autoencoder model reads the sequential data input step-by-step. The hidden state or output represents an internal learned representation of the entire input as a fixed-length vector. The decoder uses this vector as an input and interprets it as each step in the output. Unsupervised autoencoder-based Anomaly Detection techniques use the reconstruction error computed by generated output as key anomaly score identifier. The reconstruction error we use in this paper is the Mean Absolute Error (MAE).

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{1}$$

where y_i is prediction, x_i is true value, and n is the total number of data points.

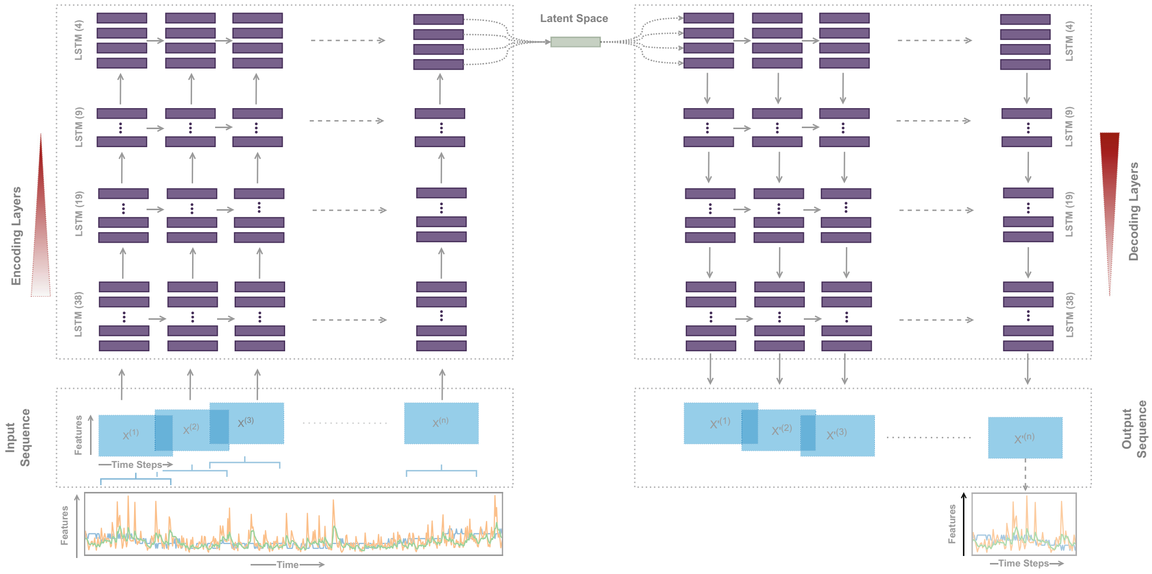

Fig. 1. Architecture of the LSTM Autoencoder used in the experiments.

3.2 Explainable AI Methods

In this section, the different available Explainable AI methods applicable to Time-Series data are described.

Additive Feature-Based. Shapley Value Sampling [10] is an additive feature attribution method based on Shapley Values. Shapley Values originate from cooperative game theory and calculate the unique and fair attribution of each feature to the output. Fair feature attribution is achieved by computing the weighted average contribution of the given features by comparing the attribution score between when the feature is present and when the feature is absent over all the possible feature subsets of each case which have an explanatory model as a linear function of binary variables:

$$g(z') = \varphi_0 + \sum \varphi_i z'_i \tag{2}$$

where $z' \in \{0,1\}^M$, M is the number of simplified input features, and $\varphi_i \in \mathbb{R}$. However, the exact computation of Shapley Values is computationally challenging for a large number of features. To that extent, Shapley Value Sampling is used to estimate the Shapley values by choosing a new random permutation of the input feature.

Perturbation-Based. Feature perturbation methods are applied to find the minimum subset of features that are enough to provide a good prediction performance by perturbing some of the original input features with some noise values. The assumption here is that if a feature is irrelevant, it can be expected to have little impact on the model performance. Thus, ablating the irrelevant features, permutation methods find a minimum subset of features that are significant for the model to issue a prediction. Feature Ablation is a method that replaces each input feature with a reference value and computes the difference in feature attribution. Similar to the Feature Ablation [2], Feature Occlusion [22] compares the group of contiguous features with reference values [12]. [16] provides a detailed study about Occlusion-based explanations in deep recurrent models for biomedical signals. Another method is Feature Permutation [2] which involves comparing the difference between individual permutation of features within a batch with the shuffled outputs of that batch.

Gradient-Based. Gradient-based methods are formulated around the similar assumption used in perturbation-based methods. The main difference is that salient features are found through propagating activation differences with respect to local variability of the features along the path from a different reference value to input. In order to capture the information flow through the network better, several approaches have been proposed with different ways to compute gradients that propagate more quantitative information than direct gradients. The Integrated Gradient is one of the prevalent methods which computes feature importance by approximating the integral of gradients of the model's output using Riemann summation [21]. The integrated gradient along the i^{th} dimension for an input x and reference value x_i is defined as follows.

$$IG_i\left(x\right) := \left(x_i - x'_i\right) \times \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha \times \left(x - x'\right)\right)}{\partial x_i} \tag{3}$$

The other method we use is Gradient Shap [10]. Gradient Shap adds Gaussian noise to select a random point along the path from reference data to input and then computes the expectation of gradient of model's prediction with respect to the randomly selected point through the additive composition of features.

Attention-Based. By adding an attention layer into the neural network architecture, we can compute the importance of features explicitly as part of training [5]. We applied the above-mentioned approaches from saliency methods on LSTM Autoencoder. The details of our experiments, as well as the methodology used, will be discussed in the following sections.

4 Proposed Solution

While most state-of-the-art anomaly localization approaches are either depending on the anomaly detection pipeline or describe the probability of features contributions by statistical study, we show for the first time that Saliency Explainable AI can be utilized in multivariate time-series Anomaly localization applications. An overview of proposed Multi-sensor Anomaly Localization Pipeline is shown in Fig. 2.

In this study, we experimentally evaluate Anomaly Localization by Saliency Explainable-AI approaches. Saliency methods were originally utilized in image classification, and there have been few attentions of using these approaches in time-series studies [5]. In our experiments, we use the Mean Absolute Error Eq. 1 in the wrapper function. To that extent, we focus on features contributing to the reconstruction error rather than reducing it. In this work we are utilizing the notion of interpretabilty of Explainable AI based on the categorizations in [26] and the numerical output [25] is accounted for in the design and execution of the experiments.

The approach to verify an explanation method is implemented in three main phases: training the model, defining ta wrapper function to interface with Explainable AI (XAI) methods, and evaluating the output of XAI approaches. In the first step, we train a black-box model on temporal data of any dimension to represent the normal behaviour of the data. Next, the trained model is used to generate new samples from test inputs. Later we construct a wrapper to integrate an autoencoder into saliency Explainable AI methods.

The purpose of the saliency methods is to determine the features or sensors that contribute to the prediction issued by a model. However, autoencoder methods learn and reconstruct the normal behaviour of the temporal data. Hence, unlike other autoencoder settings (e.g., image segmentation), there is no direct accuracy score attributed to the output in rare-event detection. For that, we trace the input to the reconstructed features via the reconstruction loss so that the explanation model signifies the features that impact the reconstruction error. Lastly, we compare features with positive contribution values with the ground truth features over the whole anomalous regions to verify how accurate the explanation methods signify the sensor localization.

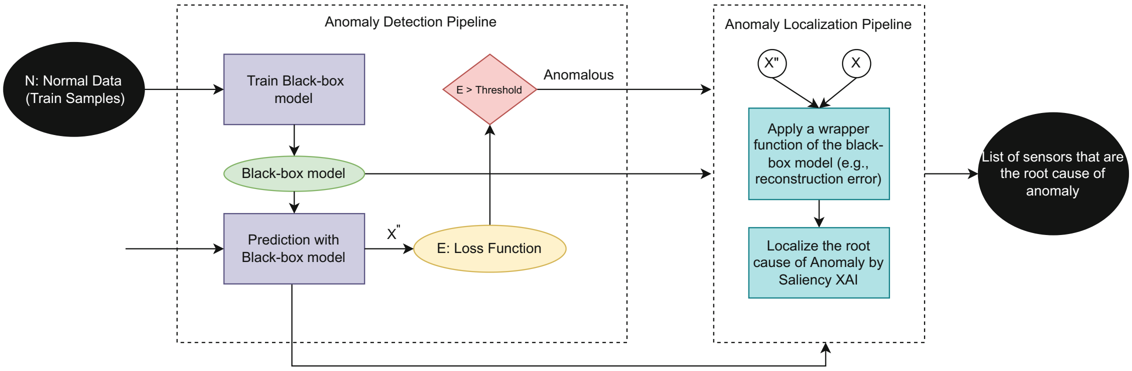
Anomaly Detection Pipeline
N: Normal Data (Train Samples)
Train Black-box model
E > Threshold
Anomalous
Black-box model
Prediction with Black-box model
X''
E: Loss Function
Anomaly Localization Pipeline
X''
X
Apply a wrapper function of the black-box model (e.g., reconstruction error)
Localize the root cause of Anomaly by Saliency XAI
List of sensors that are the root cause of anomaly

Fig. 2. Overview of the proposed Multi-sensor Anomaly Localization Pipeline.

5 Experiments

This section applies the explanation methods described in Sect. 3 to a black-box model and evaluates the quality of the feature attribution for each method on temporal anomalies. We inspect the quality of the XAI methods on local anomalous regions by comparing the most relevant sensors with the annotated sensors. This is achieved by calculating the average feature attribution score over the true anomaly regions, and comparing features sorted by attribution scores with the true anomalous sensors.

5.1 Dataset

In the experiments we used Server Machine Dataset (SMD) which is collected by [20]. SMD is a 5-week-long multivariate dataset collected from a large Internet company. It is comprised of data from 28 subsets, each being collected from different machines. For each subset, the dataset is divided into two parts of equal length of training and testing datasets. Each subset has 38 features such as CPU load, network usage, memory usage, etc. In addition to anomalies being labeled, features contributing to the anomaly at each anomalous region are also annotated, making it possible to evaluate feature attributions. The number and indices of features inducing anomalies vary upon each anomalous region, and the annotated features do not contain feature attribution values.

5.2 Experiment Setting

Explainable AI methods for the purpose of anomaly localization could be utilized to interpret any black-box model, the validation technique is model-agnostic and can be implemented using any model that is desired for multivariate temporal data analysis, such as TadGAN [7], LSTM [8], DAGMM [28], and OmniAnomaly [20]. Nevertheless, it is important to mention that the performance of any explanation model for computing the feature attribution depends on the quality of the black-box model.

In our experiments, we utilize a simple bottleneck LSTM autoencoder consisting of a two-layer encoder and decoder as a baseline model. Training the model is unsupervised and we considered the first part of the data as normal data. Firstly, features are encoded and then the latent feature representation is reversed in the same order in the decoder part to reconstruct the initial data. The increase in reconstruction error for anomalous data will lead us to detect anomalies. During the preprocessing, the data is normalized by min-max scaling, and then it is segmented into sequences through a sliding window of length 50.

We use the ADAM optimizer and stochastic gradient descent with a learning rate of 10^{-3} with a mini-batch size of 64 to train the model. We train the model with an early stopping technique on the validation data back from 20% of the training. Furthermore, to deal with the gradient overflow and prevent the exploding gradients, we incorporate a gradient clipping of norm with 5.0 as a limit.

5.3 Evaluation Metrics

Since the attribution value of true salient features is unknown, we measure how most relevant features are represented in the true saliency list for the given anomalous regions. To calculate this, we use Intersection over Union (Jaccard) similarity, Sørensen–Dice (Dice) similarity, and simple Accuracy coefficients to compute the relevance between the annotated features with size n and features with positive attribution score, ordered by their anomaly contributions. The Sørensen–Dice similarity metric is the same as the F1 score and for better representation, we use F1 Score terminology. For a ground truth y and positive attribution set y', ordered by their anomaly contributions, where K is the number of the true anomalous features, the IoU, F1, and Acc similarity coefficients are given as follows:

$$IoU\left(y',y\right)=\frac{1}{K}\sum\left(\frac{|\{y'\wedge y\}|}{|\{y'\vee y\}|}\right) \tag{4}$$

$$F_1\left(y',y\right)=\frac{1}{K}\sum\left(\frac{2\times|\{y'\wedge y\}|}{|\{y'\vee y\}|+|\{y'\wedge y\}|}\right) \tag{5}$$

$$Acc\left(y',y\right)=\frac{1}{K}\sum\left(\frac{|\{y'\wedge y\}|}{|y'|}\right) \tag{6}$$

For example, for a 7-dimensional observation x_t, a feature attribution score FA_t of {“3”:1.5, “1”:0.3, “7”:0.23, “6”:0.2, “4”:0.1, “5”:−0.3, “2”:−0.35} and ground truth GT_t of {“1”, “2”, “6”, “7”}, the evaluation results are $IoU = 0.5$, $F1 = 0.66$, and $Acc = 0.75$.

In some cases, the number of explained features with a positive attribution scores may be larger than the number of true annotated features and sometimes vice-versa as the number of features with the reducing impact on the anomaly is more. The latter happens when only a few numbers of features show enough deviation to propagate through the network and perturb all the outputs that result in high reconstruction error for the given model. The performance of the suggested evaluation metrics depends on the number of elements picked out from the predicted attribution list. Hence, the way in which we combine the attribution result with the ground truth data affects the overall assessment.

To evaluate the interpretability of anomalous regions, we compare real anomalous features data with that of predicted anomalous features. For instance, one of the direct ways to determine which fraction of the attributed futures to use for the comparison is to pick the top n features with the highest attribution score. The number n here is the number of ground truth feature that account for the anomaly. However, this method alone fails to give us sensible reasoning about the capability of the explanation methods. Therefore, we conducted slightly different experiments motivated by the top-k hit ratio which is applied for the recommendation systems [27] to achieve a more extensive and robust evaluation of the performance of the explanation methods. Hence, we employ three different ways to pick out the different numbers of top-k relevant features from the feature attribution list for comparison with the annotated features. The first approach is GT@%120, where instead of selecting the top n features with the highest positive attribution score from the attribution list, we choose the top $n * 120\%$ features.

The purpose is to give the above-mentioned evaluation metrics more flexibility as sometimes some features located close to the top n features in the attribution list are also found in the ground truth list. In the second experiment, FA@%80, we only single out the features that account for 80% of the overall attribution score to understand how well the features with the largest attribution score are represented in the ground truth list. With FA@5, we only use the top 5 features from the calculated attribution list to examine the relevance of the explanation methods regarding human reasoning, as humans inherently are only able to make reasoning about small chunks of information.

6 Results

Finally, we test the saliency Explanation AI methods by connecting them to the reconstruction loss using the wrapper function. We report in Table 1 the localization performance of different saliency Explainable AI approaches over the wrapper of reconstruction error of the LSTM Autoencoder. Moreover, to examine whether the proposed explanation methods are effective, we conduct a baseline evaluation for the LSTM Autoencoder model. In this baseline experiment, we compare the annotated features directly to that of individual reconstruction, scores ordered by their reconstruction error, using the same evaluation metrics motioned in Sect. 5.3. Furthermore, we examined OmniAnomaly [20] localization, which provides interpretations based on the reconstruction probabilities of its constituent univariate time-series with our evaluation metrics to conduct a comparison.

The overall findings in our experiment setting show that all the explanation methods, especially Occlusion, Kernel Shap, and Integrated Gradients methods

Table 1. Localization performance of different saliency Explainable AI approaches over the wrapper of reconstruction error of LSTM Autoencoder.

Approach	GT@120%			AT@80%			AT@5		
	IoU	F1	Acc	IoU	F1	Acc	IoU	F1	Acc
Baseline	43.66	52.45	50.92	32.76	43.24	45.72	37.67	46.84	57.5
OmniAnomaly	44.96	59.77	62.5	42.45	58.17	**70.75**	35.19	50.68	48.22
Occlusion	**64.39**	**76.66**	**73.66**	**62.44**	**75.44**	66.01	**47.04**	**59.18**	**72.5**
Kernel Shap	55.71 ± 0.64	66.50 ± 0.68	64.60 ± 0.76	39.07 ± 0.39	50.04 ± 0.54	66.87 ± 0.51	42.06	53.09	65.0
Shapley Val. Samp.	50.27 ± 1.29	59.68 ± 1.51	60.80 ± 1.61	39.16 ± 0.62	49.73 ± 0.83	46.61 ± 0.97	38.41 ± 1.36	47.87 ± 1.94	48.40 ± 1.53
Gradient Shap.	51.92 ± 2.41	60.18 ± 1.51	60.57 ± 1.51	30.46 ± 0.52	41.37 ± 0.68	60.63 ± 0.66	38.16	47.61	57.5
Integrated Grad.	55.61	65.65	64.26	38.54	49.52	66.25	42.06	53.09	65.0

perform significantly better than the baseline experiment, which supports the positive effect of the explanation methods for anomalous feature localization. We can see that the methods based on the Shapley value estimation yield very similar results in feature explanation. Kernel Shap method, which uses a more elaborate weighted linear regression to estimate the Shapley value, performs better than the rest of the Shapley-based solutions used in the experiment. Nevertheless, these methods indicate some variation that might induce lower stability in practice. This is caused as Kernel Shap, Shapley Value Sampling, and Gradient Shap methods use different estimation techniques to approximate the Shapley values rather than directly calculating them. The same variation is not observed in Ablation and Integrated Gradient methods as the direct solutions are possible for both.

Furthermore, explanation methods based on Ablation, particularly Occlusion, are more robust in inducing anomalous features in multivariate time-series as it outperforms all the other methods significantly for our test case. Though the Occlusion method is not the as sophisticated a method as the remaining ones, the reason that it yields better results might be due to the simple and direct connection between the occluded values (based on good reference values from good training data without any unknown anomaly) and the wrapper function.

We can observe that this result holds the same across all the different top-k ratio experiments. Though the results of the AT@%80 are less relevant than that of GT@%120, the gap between is being not too wide suggests that the features that account for 80% or more fraction of the overall attribution score are mainly rare-event inducing features. The result of *Acc* metric in FA@5, which uses only the attributed feature list as a denominator, demonstrates that the explanation methods such as Occlusion, Integrated Gradient, and Kernel Shap still manage to show significant interpretation ability compared to the baseline method when an only small fraction of feature attribution list is selected. This proves that some of the explanation methods can give considerably interpretation for domain experts in practice when a small number of features with the highest attribution score are selected for self reasoning.

The second baseline method we use is OmniAnomaly localization [20]. Although OmniAnomaly feature localization based on the reconstruction values outperforms the result of that of the baseline based on reconstruction values

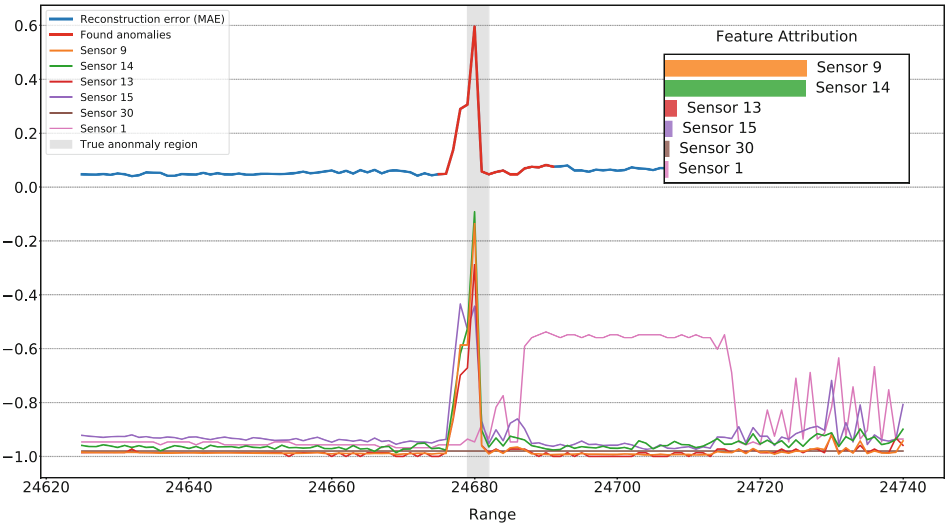

Fig. 3. An example of multi-sensor anomaly detection and localization for the given range between 24675 and 24691. The list of sensors depicted in the figure accounts for 99% of the total attribution score estimated by the Kernel Shap method for the selected range. The sensor attribution result is given in the top-right corner of the figure. The actual salient sensors for that range are 9, 13, 14, and 15^{th}.

from LSTM Autoencoder, it does perform poorer than XAI methods in most of the experiments. This proposes that the explanation methods improve the quality of feature interpretation to a greater extent than the baseline performance conducted even with a better autoencoder architecture.

In Fig. 3, an example of Multi-sensor anomaly detection and Localization for the given range between 24675 and 24691 is demonstrated. The list of sensors depicted in the figure accounts for 99% of the total attribution score estimated by the Kernel Shap method for the selected range. Note that the values of these features are not in their original state as a representation of their scaled values. The sensor attribution result is given in the top-right corner of the figure. The actual salient sensors for that range are 9, 13, 14, and 15 which are the root cause of anomaly in this specific detected anomaly.

7 Conclusion

In this work, we addressed the issue of anomaly localization in an unsupervised multivariate setup. In particular, we focused on the interpretability of LSTM Autoencoders which is one of the state-of-the-art algorithms in Multivariate time-series anomaly detection. To this end, we have applied saliency Explained AI approaches. These approaches could be applied over different families of networks for anomaly detection. As a baseline, we tested these approaches on LSTM Autoencoder and SMD datasets. The explanation could be utilized in unsupervised multivariate anomaly localization. Previous work in Explainable AI has often looked at how to interpret different AI models and contrast their results.

Going forward, we think it is crucial to also investigate applications of Explainable AI. This might include studying the impact of explanations on trusting AI solutions and bridging the gap between AI practitioners and domain experts.

Acknowledgments. Resources used in preparing this research have received funding by the Federal Ministry for Economic Affairs and Climate Action (BMWK) in the research project SPAICER. The collaboration is among partners of the project: German Research Center for Artificial Intelligence (DFKI) and TU Darmstadt. The proposed approach is going to be part of Smart Resilience Services.

References

1. Antwarg, L., Miller, R.M., Shapira, B., Rokach, L.: Explaining anomalies detected by autoencoders using shapley additive explanations. Expert Syst. Appl. **186**, 115736 (2021)
2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
3. Carletti, M., Masiero, C., Beghi, A., Susto, G.A.: Explainable machine learning in industry 4.0: evaluating feature importance in anomaly detection to enable root cause analysis. In: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 21–26 (2019)
4. Choi, Y., Lim, H., Choi, H., Kim, I.J.: Gan-based anomaly detection and localization of multivariate time series data for power plant. In: 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 71–74 (2020)
5. Crabbe, J., van der Schaar, M.: Explaining time series predictions with dynamic masks. In: ICML (2021)
6. Fisher, A.J., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. JMLR **20**, 1–81 (2019)
7. Geiger, A., Liu, D., Alnegheimish, S., Cuesta-Infante, A., Veeramachaneni, K.: Tadgan: time series anomaly detection using generative adversarial networks. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 33–43 (2020)
8. Hundman, K., Constantinou, V., Laporte, C., Colwell, I., Söderström, T.: Detecting spacecraft anomalies using LSTMS and nonparametric dynamic thresholding. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2018)
9. Jiang, R., Fei, H., Huan, J.: Anomaly localization for network data streams with graph joint sparse PCA. In: KDD (2011)
10. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. arXiv arXiv:abs/1705.07874 (2017)
11. Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., Shroff, G.: LSTM-based encoder-decoder for multi-sensor anomaly detection. arXiv preprint arXiv:1607.00148 (2016)
12. Meyes, R., Lu, M., de Puiseau, C.W., Meisen, T.: Ablation studies in artificial neural networks. arXiv:abs/1901.08644 (2019)
13. Mozaffari, M., Yılmaz, Y.: Multivariate and online anomaly detection and localization for high-dimensional systems (2019)
14. Mujkanovic, F., Doskoc, V., Schirneck, M., Schäfer, P., Friedrich, T.: Timexplain - a framework for explaining the predictions of time series classifiers. arXiv:abs/2007.07606 (2020)

15. Pan, Q., Hu, W., Zhu, J.: Series saliency: temporal interpretation for multivariate time series forecasting. arXiv abs/2012.09324 (2020)
16. Resta, M., Monreale, A., Bacciu, D.: Occlusion-based explanations in deep recurrent models for biomedical signals. Entropy **23**, 1064 (2021)
17. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
18. Roelofs, C.M., Lutz, M.A., Faulstich, S., Vogt, S.: Autoencoder-based anomaly root cause analysis for wind turbines (2021)
19. Shankaranarayana, S.M., Runje, D.: Alime: autoencoder based approach for local interpretability. arXiv:abs/1909.02437 (2019)
20. Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., Pei, D.: Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019)
21. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. arXiv:abs/1703.01365 (2017)
22. Suresh, H., Hunt, N., Johnson, A.E.W., Celi, L.A., Szolovits, P., Ghassemi, M.: Clinical intervention prediction and understanding with deep neural networks. In: MLHC (2017)
23. Tonekaboni, S., Joshi, S., Campbell, K., Duvenaud, D.K., Goldenberg, A.: What went wrong and when? Instance-wise feature importance for time-series black-box models. In: NeurIPS (2020)
24. Trifunov, V.T., Shadaydeh, M., Barz, B., Denzler, J.: Anomaly attribution of multivariate time series using counterfactual reasoning. In: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 166–172 (2021)
25. Vilone, G., Longo, L.: Classification of explainable artificial intelligence methods through their output formats. Mach. Learn. Knowl. Extr. **3**(3), 615–661 (2021)
26. Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion **76**, 89–106 (2021)
27. Yang, X., Steck, H., Guo, Y., Liu, Y.: On top-k recommendation using social networks. In: Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys 2012, pp. 67–74. Association for Computing Machinery, New York (2012). https://doi.org/10.1145/2365952.2365969
28. Zong, B., et al.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: ICLR (2018)