In: 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC),
24-27 September 2024, Edmonton, Canada, pp. 3265-3270
DOI:10.1109/itsc58415.2024.10920008

Enhancing Performance and Interpretability of Multivariate Time-Series Model Through Sparse Saliency

Xiangqi Kong[1], Yang Xing[1*], *Senior Member, IEEE,* Zeyu Liu[1], Antonios Tsourdos[1*], and Andreas Wikander[2]

***Abstract*— Explainable time-series modelling is an essential task for modern intelligent transportation systems (ITS). However, balancing accuracy and interpretability in multivariate time series forecasting presents significant challenges. These challenges arise from the necessity to understand the significance of features and their temporal variations. Factors such as autocorrelation in time series and data processing techniques like sliding windows expand feature sets, thereby complicating pattern recognition using traditional post-hoc explanation methods and making the issue even more complex. To overcome these challenges, in this study, we propose a flexible post-process approach which generates sparse and normalized saliency values based on existing saliency generation methods such as GradientSHAP. Additionally, an optional window aggregation and alignment strategy is introduced to align with the original time series dataset, enhancing the intuitive understanding of feature importance. Furthermore, the potential use of sparse saliency for data augmentation to improve the model is explored. Lastly, we utilize naturalistic data from San Francisco airport to demonstrate our approach for ITS time-series prediction and explanation. The evaluation results indicate that integrating sparse saliency from high-performing models not only boosts the performance of XGBoost models by 10.92% but also simplifies model complexity, facilitating easier interpretation.**

***Index Terms*— Interpretable multivariate time series modelling, Saliency generation techniques, Post-hoc explanation**

I. INTRODUCTION

A. Motivation

Multivariate time series forecasting plays a crucial role in intelligent transportation systems (ITS), which includes domains such as traffic management, autonomous driving, and fault diagnostics [1]. Traffic prediction plays a pivotal role in traffic planning and network optimization, enabling the anticipation of future trends and states of interconnected variables critical for informed decision-making and strategic planning. Moreover, the analysis of large-scale time series data from ITS often necessitates human intervention [2], either in real-time or post-hoc, to adhere to ethical and regulatory policies [3]. To mitigate the decision-making risks inherent in black-box models such as Transformers and recurrent neural networks, researchers are focusing on two primary approaches: enhancing model transparency [4] and utilizing post-hoc explanations to interpret model predictions [5].

Among the post-hoc techniques, extracting and analyzing feature importance through saliency values have gained prominence [6][7][8]. However, the understandability of these techniques heavily depends on the performance of the underlying models and the explanation methods used [9][10][11]. Common attribution methods like GradientSHAP [12] often struggle with interpreting time series data, producing dense and noisy outputs that obscure relevant information, especially when the data has underlying temporal or latent patterns [13]. Furthermore, the rising dimensionality in multivariate time series introduces complexities due to autocorrelation and the application of data processing techniques, such as sliding window methods. These factors complicate the extraction of meaningful insights and make it difficult for humans to derive clear understandings from saliency maps.

In response to these challenges, this study introduces a method that enhances current saliency techniques by integrating sparse and normalization strategies, thereby reducing complexity and minimizing minor disturbances. Additionally, it aggregates fine-grained saliency maps over meaningful time ranges, contributing to more comprehensible results for human users. Furthermore, we explore whether sparse saliency values from advanced models can enhance the predictive capabilities of simpler models like XGBoost, which are commonly used in multivariate time series analysis but may underperform compared to more sophisticated neural network architectures.

B. Contribution

This study makes several key contributions:

- Human-readable Saliency Maps: we develop a method that refines existing saliency generation techniques to produce sparser saliency maps. This simplification aids in clearer pattern recognition, could potentially reduce the cognitive load on analysts for end-users.
- Performance Enhancement Validation: we empirically validate the effectiveness of sparse saliency in improving the performance of underperforming models. The final results demonstrate the efficiency of the proposed sparse saliency enhancement methodology.
- Reduction of Model Complexity: our experiments show that integrating sparse saliency into the learning process can reduce the overall complexity of the XGBoost model while simultaneously enhancing its performance.

This project is supported by the Engineering and Physical Sciences Research Council (EPSRC) training grant entitled "DTP 2020-2021 Cranfield University" bearing reference EP/T518104/1.

[1]Xiangqi Kong, Yang Xing, Zeyu Liu, and Antonios Tsourdos are with the School of Aerospace, Transport and Manufacturing, Cranfield University, Bedfordshire, MK43 0AL, U.K.

[2]Andreas Wikander is with Saab, 417 56 Göteborg, Sweden.

*Corresponding authors: Yang Xing and Antonios Tsourdos (email: yang.x@cranfield.ac.uk, a.tsourdos@cranfield.ac.uk)

C. Paper Organization

The paper is structed as follows: Section II reviews related works, highlighting the advancements of time series modelling, saliency generation and saliency guided learning. Section III details the proposed methodologies, including the theoretical underpinnings and practical implementations of the sparse saliency techniques. The experimental setup and results analysis are presented in Section IV. Finally, the study concludes in Section V with a summary of findings, limitations and potential directions for future research.

II. Related Works

A. Time series modelling

Researchers have proposed multivariate models to capture the complex interactions and correlations among different traffic variables. One such model is SARIMAX [14], which incorporates seasonal autoregressive integrated moving average principles to account for seasonal patterns. However, traditional prediction methods may not be effective in situations where external exceptional events, such as pandemics, introduce significant uncertainty in forecasting. To address these challenges, machine learning models have been employed. LightGBM has shown superior performance compared to baseline models that rely solely on historical passenger data, time-related inputs, and quarterly Gross Domestic Product (GDP) by considering epidemic-related variables to predict future air passenger demand [15]. In bus passenger flow prediction tasks that integrate point-of-interest data, XGBoost has shown better training efficiency and prediction accuracy compared to other models [16].

Deep recurrent neural networks like long short-term Memory (LSTM) capture long-term dependencies in sequential data [17]. Gated recurrent units (GRU) simplifies the architecture by update and reset gates, making it computationally more efficient while still effectively capturing temporal dependencies [18]. For large datasets and long forecasting horizons, transformer-based models have proven to be effective due to their ability to model complex dependencies in time series data. The temporal fusion transformer (TFT) integrates static and time-varying data to capture global and local temporal patterns, enhancing forecasting accuracy and interpretability with attention mechanism [4]. The Autoformer improves upon traditional Transformer architectures by integrating an automatic decomposition mechanism [19]. In addition, the Timexer model excels in time series forecasting with exogenous variables, especially in handling low-quality exogenous data and capturing inflection points in the target variable. This model supports multi-level prediction and can forecast multiple future time steps. Techniques like inverted embedding and patch embedding are incorporated to enhance its performance [20]. Constructing auxiliary time series from the original time series, which incorporates inter-series relationships also improve performance. This approach, combined with a simple 2-layer MLP as the core predictor, significantly reduces complexity and parameter count [21].

B. Saliency methods for time series

Saliency generation in time series data can be categorized into three main methods: gradient-based, perturbation-based, and attention-based. Gradient-based methods compute gradients of the model's output with respect to the input time series, using these gradients to create saliency maps that highlight influential segments or features. Attention mechanisms, exemplified by the self-attention in Transformers, generate saliency maps by computing attention weights for each timestep or feature, indicating their importance for model predictions. However, traditional saliency techniques like Grad-CAM [22], SHAP [12], originally designed for image and natural language processing, do not naturally suit time series data. These methods often fail in tasks where the detection of latent patterns—such as dominant frequencies or trend dependencies—is essential, which is common in applications like physiological signal analysis and fault diagnosis. Evaluations on time series classification tasks reveal that existing perturbation and gradient-based methods sometimes lack consistency and robustness [10][13].

Recently, several saliency generation methods specific to time series have been proposed, Series saliency method extracts "series images" from sliding windows of the time series and applies a saliency map segmentation based on the smallest destroying region principle. This approach considers both the feature and temporal dimensions, improving accuracy and interpretability [23]. Jonathan Crabbé et al. introduce Dynamic Masks, which employ perturbation masks tailored to time series data to capture temporal dependencies and identify relevant features for predictions. These masks are optimized for parsimony and legibility, improving feature importance identification [8]. Joseph Enguehard [24] proposes learning perturbations to explain time series predictions, where a trainable mask is used to perturb the input and identify influential features and their timing. This approach learns both the masks and associated perturbations, resulting in enhanced explanations. Additionally, the Time Interpret library extends Captum for temporal data explanation, providing saliency generation methods and evaluation tools [25]. The ContraLSP framework combines contrastive learning and sparse gate techniques, introduces counterfactual samples and maintains the distribution of the data, resulting in improved explanation quality and alleviating distribution shift issues [26].

The core idea behind saliency guided training is to focus the model's attention on the most prominent regions of the input during the training process, instead of treating all parts equally [27]. The saliency guided training procedure involves iteratively masking features that have small and possibly noisy gradients, while also aiming to maximize the similarity of model outputs for both the masked and unmasked inputs. This approach has shown effectiveness in various domains, including images, language, and multivariate time series [28].

III. Methodology

A. Overall Framework

As depicted in Fig. 1, we propose a novel framework that incorporates sparse saliency maps to enhance interpretability

and accuracy in time series prediction models. The process begins with dataset preparation, where the raw dataset D undergoes a windowing technique to create sequences that capture temporal dependencies. The variable *seq_len* represents the number of time steps or observations included in each window created from D. The window is expressed as:

$$\mathbf{W}_i = \mathbf{D}[i : i + seq\_len] \tag{1}$$

Following data preparation, we obtain D'_{train}, multiple forecasters $\{M_i\}$ are then trained using D'_{train}, and the optimal model M is selected based on performance metrics. An explainer subsequently generates the original saliency maps A from M.

The sparse saliency generation phase processes A through a proposed sparse and normalize function E, resulting in sparse saliency maps A' that emphasize the most impactful features. The data dimensions are then aggregated from $(W \times T \times F)$ to $(T \times F)$. The approach of utilizing post-processing to introduce sparsity provides the framework with enhanced flexibility.

In this study, for dataset augmentation part, the sparse saliency values A' are simply integrated with the corresponding positions of the raw training subset D_{train} using element-wise multiplication, denoted as:

$$\mathbf{D}''_{\text{train}} = \mathbf{A}' \odot \mathbf{D}_{\text{train}} \tag{2}$$

This enhanced dataset serves in the final phase, model update, where M is retrained or updated with D''_{train}, resulting in an enhanced model M' that incorporates refined insights for improved performance and interpretability.

B. Sparse Normalize Saliency

This section clarifies the sparse saliency generation phase and introduces the proposed function E, designed to enhance the clarity and readability of saliency maps by integrating mask filtering with adaptive scaling techniques. Function E selectively emphasizes the most impactful features, thereby producing sparser and more interpretable visual representations. This improves the understanding of model decisions, particularly in complex datasets.

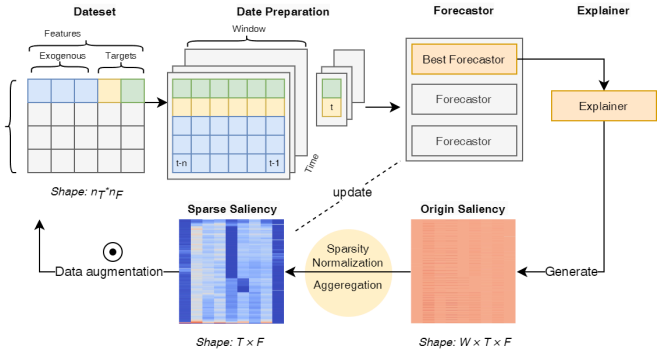

Fig. 1. Illustration of the proposed framework, integrating saliency maps into the forecasting model lifecycle. This diagram highlights the process of normalizing, and sparsifying saliency maps to enhance predictive accuracy and interpretability.

Given a tensor of original saliency values A generated by an explainer, which signifies the importance assigned to each feature by a model, the function aims to normalize these attributions using scaling and thresholding mechanisms controlled by the parameters percentile p, scalar α, and normalization method.

First, we define a threshold T based on the given percentile p of the absolute values of attributions A:

$$T = \ quantile\ \left(|A|, \frac{p}{100}\right) \tag{3}$$

The scale factor S is computed using the standard deviation σ of the attributions A and the given scalar α, which influences how much the standard deviation of the attributions affects the scaling of the data:

$$S = \frac{\alpha}{\sigma(A)} \tag{4}$$

A mask $\mathbb{M}$ is created to identify elements in A that are greater than or equal to the threshold T:

$$\mathbb{M} = |A| \geq T \tag{5}$$

Depending on the chosen method $\mathscr{M}$, scaled attributions A' are calculated. For the hyperbolic tangent method ($\mathscr{M}$=*tanh*):

$$A'[i] = \begin{cases} \frac{e^{S \cdot A[i]} - e^{-S \cdot A[i]}}{e^{S \cdot A[i]} + e^{-S \cdot A[i]}} & \text{if } \mathbb{M}[i] = \text{True} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

When $\mathscr{M}$=*Softmax*:

$$A'[i] = \begin{cases} \frac{\exp(S \cdot A[i])}{\sum_{j \in J} \exp(S \cdot A[j])} & \text{if } \mathbb{M}[i] = \text{True} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where J is the set of indices where $\mathbb{M}$ is True.

To represent the function E in a formalized manner, we can summarize and combine the steps into a sequence of operations. The entire process can be expressed in a formula:

$$E(A, p, \alpha, \mathscr{M}) = A' \tag{8}$$

In summary, by applying mask $\mathbb{M}$ and normalization operations using the *tanh* or *softmax* methods, the sparsity of the original saliency map derived from the model M is increased while the redundant information from dense attributions is reduced. Consequently, the salient features become more prominent, facilitating human understanding and enhancing the model's interpretability in practical applications.

IV. Experiments and Analysis

A. Experiment Setup

This section details the experimental setup, designed as follows:

1) Data preparation: In our study, we used the Airport Passenger Flow Dataset obtained from San Francisco Airport [29] as our primary data source. This dataset allowed us to analyze the monthly passenger flow across eight different boarding areas (A-G, Other). To forecast future traffic for a specific subsequent month, we segmented the data into 12-month windows and utilized an 85% training and 15% testing split. To provide a comprehensive understanding of the variations in passenger flow influenced by external factors, we also incorporated pandemic-related search trends from Google Trends [30] as an exogenous variable, enriching the dataset. In addition, when incorporating sparse saliency values, we not only used element-wise multiplication for sparse saliency value A' with the training set D_{train} but also applied scaling using the median value of A' to the entire testing set. This approach allowed us to effectively integrate the sparse saliency values into our analysis, ensuring consistency in the scaling process across both the training and testing datasets.

2) Saliency Values Integration: We employed a suite of basic forecasters tailored to multivariate time series analysis, including XGBoost, LSTM, GRU, and Transformer models. The model delivering the best performance was then used to generate original saliency values using GradientSHAP. Sparse normalization techniques were applied with a mask percentile P of 0, 5, 10, 50, 90, a scalar α of 1, and a hyperbolic tangent activation function. The training dataset was augmented by element-wise multiplication with the sparsely normalized saliency values, whereas the testing set adjustments were based on the dataset's median saliency values. This approach confined the saliency generation process to the training data, preventing data leakage and its potential impact on the results. Then, the modified datasets were utilized for model update. The effects of varying mask percentiles on the XGBoost model's performance were analyzed, along with the performance variability across different estimators, which serve as indicators of the model's complexity. Based on our preliminary empirical tests, we recommend using the tanh function when there are many extremely low saliency values, as it converges rapidly near zero and effectively filters out more noise.

3) Measurement metrics: The effectiveness of the forecasting models is quantitatively assessed using four standard error metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Symmetric Mean Absolute Percentage Error (sMAPE), and Coefficient of Determination (R2). These metrics provide a comprehensive evaluation of the models' accuracy in forecasting future values. Considering that different saliency values may introduce varying scaling factors to the augmented training dataset in data preparation process, we take into account the need for consistent model evaluation across different scenarios. Therefore, we descale all values to their original scale before conducting the comparative analysis. This approach allows us to make fair and accurate comparisons among the different models, regardless of the variations introduced by the saliency values.

B. Results Analysis

Based on the results presented in Table I, "XGBoost+" refers to using a dataset enhanced by the GRU model, without exogenous variables at a 10% mask percentile. The more powerful Transformer model does not show a significant advantage with this relatively small dataset. On the contrary, the lightweight GRU model demonstrates the best performance on the original dataset, both with and without the inclusion of exogenous variables.

TABLE I
COMPARISON OF MODELS WITH AND WITHOUT EXOGENOUS VARIABLE

	MAE	MSE	sMAPE	R2
only history data				
XGBoost	6.63E+04	9.59E+09	341.0656	0.2072
LSTM	7.68E+04	1.03E+10	337.1754	0.7716
GRU	6.96E+04	8.89E+09	321.5465	**0.8230**
Transformer	1.10E+05	2.21E+10	383.1288	0.5237
XGBoost+	**6.20E+04**	**8.21E+09**	**308.5541**	0.3485
with exogenous variables				
XGBoost	6.64E+04	9.63E+09	341.0782	0.2327
LSTM	7.52E+04	1.04E+10	339.0567	0.7872
GRU	6.73E+04	**8.01E+09**	324.1746	**0.8237**
Transformer	9.64E+04	1.63E+10	366.8931	0.5646
XGBoost+	**6.21E+04**	8.25E+09	**308.5801**	0.3486

In Fig. 2, the predictions of GRU and XGBoost models on the test dataset with exogenous variables are compared. It is evident that XGBoost initially exhibits a significant discrepancy with the actual values, particularly in boarding area D, as indicated by the red circle. By utilizing the sparse saliency enhanced dataset generated by the GRU model with exogenous variables and a mask percentile $p = 10$, the performance of the XGBoost model with same estimators as before shows a substantial improvement. This update effectively resolves the prediction issue in boarding area D, addressing the earlier discrepancy observed.

Table II presents results demonstrating the influence of different mask percentile values of sparse saliency on the performance of the XGBoost model. Notably, even when the number of XGBoost estimators was reduced to 20, optimal performance was achieved with a mask percentile of 10. This finding underscores the improvement in model performance through the use of an updated model with an appropriate mask percentile for sparse saliency. The MAE metric indicates that XGBoost+, with a score of 6.21e+04, outperformed the previous best performance achieved by GRU, which scored 6.73e+04, resulting in a 7.73% improvement for data with exogenous variables and a 10.92% improvement for data without exogenous variables. Conversely, when directly using a saliency values generated by the GradientSHAP method, the model's performance decreased.

As illustrated in Table II, an excessive mask percentile p, such as 90, results in significant information loss, negatively impacting outcomes. However, as depicted in Figure 3, these higher p potentially enhance the detection of unusual varia-

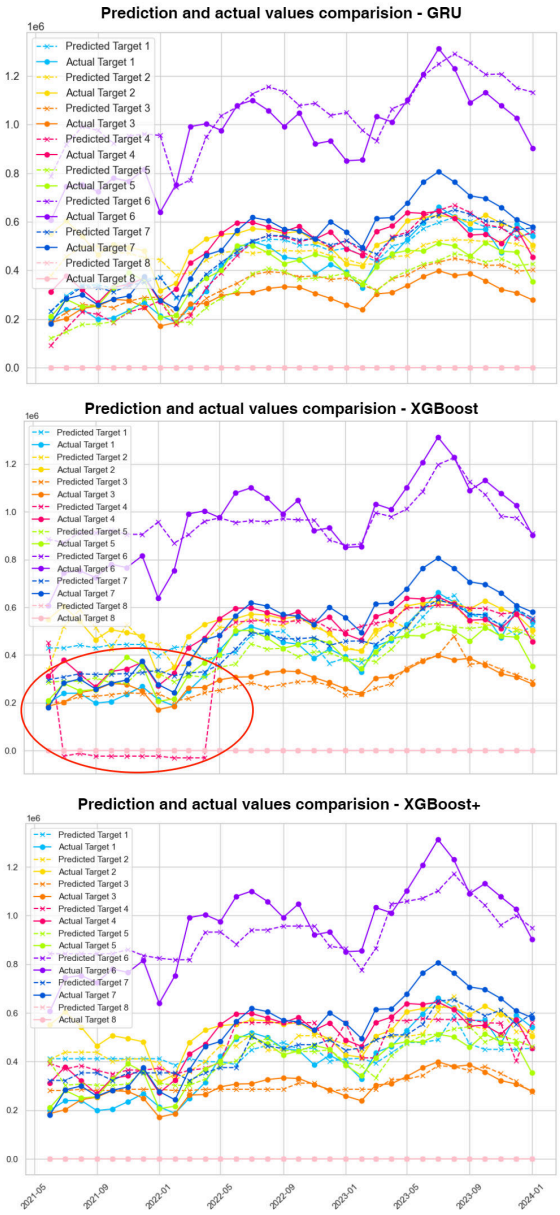

Fig. 2. Comparison of GRU, XGBoost, and XGBoost+ Predictions vs. Actual Values for Target Areas A-G and Other (labeled with target 1-8).

tions, both in individual data points and patterns. Identifying instances where the sparse saliency heatmap aligns with the ground truth becomes easier when visualizing abnormal events in the dataset.

V. CONCLUSIONS AND FUTURE WORK

This paper explores the use of post-hoc explanations to provide understandable insights into model decisions for human comprehension, improving performance by utilizing sparse saliency values generated through our proposed methodology. The study focuses on multivariate time series forecasting, demonstrating its effectiveness on real-world datasets with known ground truths.

However, the universal applicability of this approach across various time series tasks remains to be explored. Future work should include a comparative analysis of the impact of saliency values generated by different explainers after undergoing sparsification and normalization to gain further insights. Additionally, the selection of explainers and the development of improved data augmentation methods merit further investigation.

Our conclusions on readability are based on subjective judgments of visualized results, necessitating systematic evaluation on explainable interfaces [31]. Furthermore, this study utilizes a small-scale, coarse-grained time series dataset with a monthly time span, which does not demand high real-time processing. Many time series models in practical applications require real-time capabilities and are trained on finer-grained datasets. Therefore, future work should validate our findings on larger and finer-grained datasets to ensure broader applicability and performance in real-time scenarios.

TABLE II
PERFORMANCE COMPARISION FOR DIFFERENT SPARSITY SALIENCY VLAUES AND XGBOOST ESTIMATORS

mask Percentile	xgboost estimators	Descaled MAE	Descaled MSE	Descaled SMAPE	R2
0	10	6.57E+04	8.58E+09	314.0539	0.3186
0	**20**	**6.26E+04**	**8.38E+09**	**309.4975**	**0.3344**
0	40	7.01E+04	1.04E+10	320.8295	0.2339
0	60	7.21E+04	1.10E+10	324.9005	0.1864
0	80	7.26E+04	1.12E+10	325.7797	0.1690
0	100	7.32E+04	1.14E+10	326.9381	0.1548
5	10	6.57E+04	8.58E+09	314.0543	0.3186
5	**20**	**6.26E+04**	**8.38E+09**	**309.4980**	**0.3344**
5	40	7.01E+04	1.04E+10	320.8301	0.2339
5	60	7.21E+04	1.10E+10	324.9010	0.1864
5	80	7.27E+04	1.12E+10	326.0429	0.1672
5	100	7.33E+04	1.14E+10	327.0487	0.1528
10	10	6.58E+04	8.60E+09	314.1743	0.3179
10	**20**	**6.20E+04**	**8.21E+09**	**308.5541**	**0.3485**
10	40	6.99E+04	1.06E+10	320.7102	0.2303
10	60	7.19E+04	1.13E+10	324.5954	0.1647
10	80	7.28E+04	1.16E+10	326.0260	0.1459
10	100	7.39E+04	1.20E+10	327.8790	0.1233
50	**10**	**6.63E+04**	**9.03E+09**	**315.6055**	**0.2611**
50	20	6.91E+04	9.78E+09	320.2657	0.2250
50	40	7.74E+04	1.25E+10	336.0832	0.0589
50	60	8.14E+04	1.35E+10	342.9682	-0.0791
50	80	8.42E+04	1.46E+10	347.2121	-0.1596
50	100	8.51E+04	1.50E+10	347.7131	-0.1912
90	**10**	**1.03E+05**	**1.98E+10**	**368.6473**	**-0.4944**
90	20	1.12E+05	2.46E+10	387.9443	-0.7831
90	40	1.20E+05	3.22E+10	399.8461	-1.2336
90	60	1.26E+05	3.71E+10	405.6978	-1.7085
90	80	1.26E+05	3.91E+10	403.8361	-1.8246
90	100	1.26E+05	4.00E+10	403.9194	-1.9232

REFERENCES

[1] Xing, Yang, Zhongxu Hu, Peng Hang, and Chen Lv. Learning from the dark side: A parallel time series modelling framework for forecasting and fault detection on intelligent vehicles. *IEEE Transactions on Intelligent Vehicles*. (2023).

[2] Nunes, A., Reimer, B. & Coughlin, J. People must retain control of autonomous vehicles. (Nature Publishing Group UK London,2018)

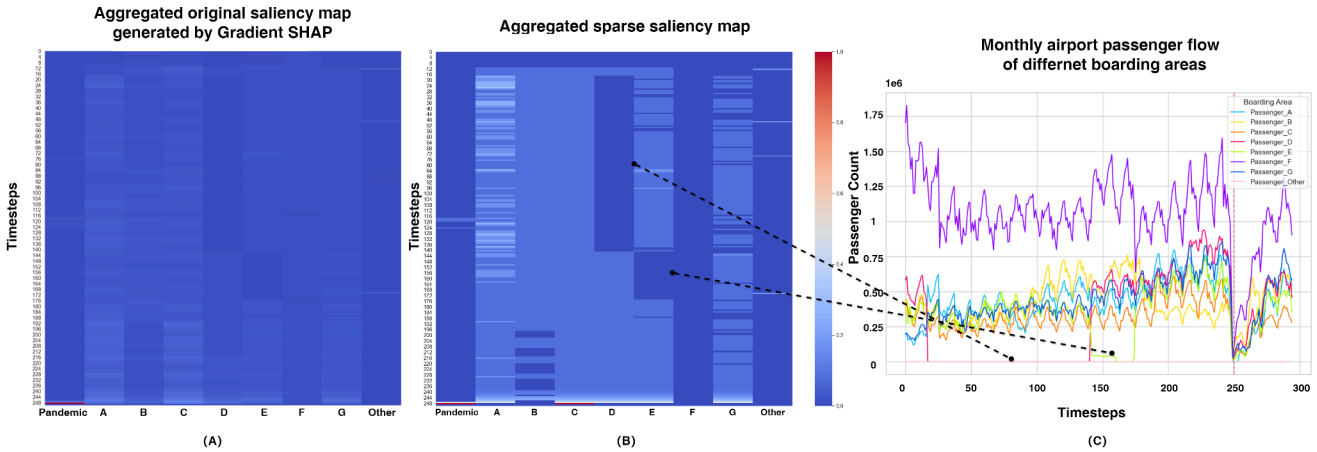

Fig. 3. Readability of Different Saliency Maps: (A) is Aggregated Original Saliency Map; (B) is Aggregated Sparse saliency Map; (C) is Visualization of Real Monthly Passenger Flow of San Francisco Airport Boarding Areas.

[3] Gaur, L. & Sahoo, B. Introduction to explainable AI and intelligent transportation. *Explainable Artificial Intelligence For Intelligent Transportation Systems: Ethics And Applications*. pp. 1-25 (2022)

[4] Lim, B., Arık, S., Loeff, N. & Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal Of Forecasting*. **37**, 1748-1764 (2021)

[5] Xie, Y., Pongsakornsathien, N., Gardi, A. & Sabatini, R. Explanation of machine-learning solutions in air-traffic management. *Aerospace*. **8**, 224 (2021)

[6] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M. & Hussain, A. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*. **16**, 45-74 (2024)

[7] Tonekaboni, S., Joshi, S., Campbell, K., Duvenaud, D. & Goldenberg, A. What went wrong and when? Instance-wise feature importance for time-series black-box models. *Advances In Neural Information Processing Systems*. **33** pp. 799-809 (2020)

[8] Crabbé, J. & Van Der Schaar, M. Explaining time series predictions with dynamic masks. *International Conference On Machine Learning*. pp. 2166-2177 (2021)

[9] Balestra, C., Li, B. & Müller, E. On the Consistency and Robustness of Saliency Explanations for Time Series Classification. *ArXiv Preprint ArXiv:2309.01457*. (2023)

[10] Turbé, H., Bjelogrlic, M., Lovis, C. & Mengaldo, G. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*. **5**, 250-260 (2023)

[11] Slack, D., Hilgard, S., Jia, E., Singh, S. & Lakkaraju, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *Proceedings Of The AAAI/ACM Conference On AI, Ethics, And Society*. pp. 180-186 (2020)

[12] Lundberg, S. & Lee, S. A unified approach to interpreting model predictions. *Advances In Neural Information Processing Systems*. **30** (2017)

[13] Schröder, M., Zamanian, A. & Ahmidi, N. Post-hoc Saliency Methods Fail to Capture Latent Feature Importance in Time Series Data. *International Workshop On Trustworthy Machine Learning For Healthcare*. pp. 106-121 (2023)

[14] Cools, M., Moons, E. & Wets, G. Investigating the variability in daily traffic counts through use of ARIMAX and SARIMAX models: assessing the effect of holidays on two site locations. *Transportation Research Record*. **2136**, 57-66 (2009)

[15] Tang, H., Yu, J., Lin, B., Geng, Y., Wang, Z., Chen, X., Yang, L., Lin, T. & Xiao, F. Airport terminal passenger forecast under the impact of COVID-19 outbreaks: A case study from China. *Journal Of Building Engineering*. **65** pp. 105740 (2023)

[16] Lv, W., Lv, Y., Ouyang, Q. & Ren, Y. A bus passenger flow prediction model fused with point-of-interest data based on extreme gradient boosting. *Applied Sciences*. **12**, 940 (2022)

[17] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Computation*. **9**, 1735-1780 (1997)

[18] Fu, R., Zhang, Z. & Li, L. Using LSTM and GRU neural network methods for traffic flow prediction. *2016 31st Youth Academic Annual Conference Of Chinese Association Of Automation (YAC)*. pp. 324-328 (2016)

[19] Wu, H., Xu, J., Wang, J. & Long, M. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. *Advances In Neural Information Processing Systems*. (2021)

[20] Wang, Y., Wu, H., Dong, J., Liu, Y., Qiu, Y., Zhang, H., Wang, J. & Long, M. TimeXer: Empowering Transformers for Time Series Forecasting with Exogenous Variables. *ArXiv Preprint ArXiv:2402.19072*. (2024)

[21] Lu, J., Han, X., Sun, Y. & Yang, S. CATS: Enhancing Multivariate Time Series Forecasting by Constructing Auxiliary Time Series as Exogenous Variables. *ArXiv Preprint ArXiv:2403.01673*. (2024)

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

[23] Pan, Q., Hu, W. & Chen, N. Two Birds with One Stone: Series Saliency for Accurate and Interpretable Multivariate Time Series Forecasting.. *IJCAI*. pp. 2884-2891 (2021)

[24] Enguehard, J. Learning perturbations to explain time series predictions. *International Conference On Machine Learning*. pp. 9329-9342 (2023)

[25] Enguehard, J. Time Interpret: a Unified Model Interpretability Library for Time Series. *ArXiv Preprint ArXiv:2306.02968*. (2023)

[26] Liu, Z., Zhang, Y., Wang, T., Wang, Z., Luo, D., Du, M., Wu, M., Wang, Y., Chen, C., Fan, L. & Others Explaining Time Series via Contrastive and Locally Sparse Perturbations. *ArXiv Preprint ArXiv:2401.08552*. (2024)

[27] Liang, A., Thomason, J. & Bıyık, E. Visarl: Visual reinforcement learning guided by human saliency. *ArXiv Preprint ArXiv:2403.10940*. (2024)

[28] Ismail, A., Corrada Bravo, H. & Feizi, S. Improving deep learning interpretability by saliency guided training. *Advances In Neural Information Processing Systems*. **34** pp. 26726-26739 (2021)

[29] City and County of San Francisco Air Traffic Passenger Statistics - DataSF. https://data.sfgov.org/Transportation/Air-Traffic-Passenger-Statistics/rkru-6vcg/about_data, [Online; accessed May-2024]

[30] Google LLC Google Trends. (https://trends.google.com/trends), [Online; accessed May-2024]

[31] Kong, Xiangqi, Yang Xing, Antonios Tsourdos, Ziyue Wang, Weisi Guo, Adolfo Perrusquia, and Andreas Wikander. Explainable Interface for Human-Autonomy Teaming: A Survey. *arXiv preprint arXiv:2405.02583*. (2024).