Response to Reviewer 1: - You are correct that "comparison with CNNs makes it convincing that the saliency problems" addressed in our paper are specific to multi-variate "timeseries models". This finding is the key message of our paper. Even though the mathematical explanation of that behavior remains an open question, we focused on identifying the problem, benchmarking that behavior across different data characteristics using objective ground truth, and proposing a rescaling approach that enhances the quality of time-series saliency methods. - Upon your suggestion, we repeat some experiments by masking features and evaluating changes in the loss (see AUPR of some datasets in the table below). We find that although Masking is very expensive, it does not perform well.

Methods	Middle Box			Rare Time		
	LSTM+In.	TCN	Trans.	LSTM+In.	TCN	Trans.
Grad	0.494	0.237	0.288	0.218	**0.334**	0.027
SG	0.427	0.312	0.295	0.025	0.134	0.012
Masking	0.332	0.348	0.382	0.180	0.000	0.028
TSR+Grad	**0.677**	**0.541**	**0.437**	**0.378**	0.103	**0.307**

Table 1: AUPR: With feature masking

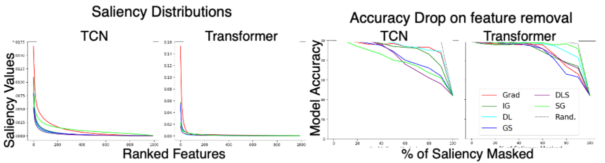

Figure 1: Experiments on new NARMA dataset

Response to Reviewer 2: Novelty and contributions. In our paper, we make following contributions: (1) We identify a limitation in the saliency methods when applied to multi-variate timeseries models, this finding itself is novel. (2) We show that this limitation exists in different neural architectures and across a large range of popular saliency methods in different setups. (3) We create benchmark metrics to evaluate models/methods that can be used on datasets where informative features are known, we show that methods should be compared based on precision and recall of features identified as salient not only with model accuracy drop upon elimination as in standard practice. (4) We propose a novel method TSR that improves the quality of saliency methods for timeseries. We believe these contributions can be of interest to the ML community.

Richness of tested datasets. The benchmark itself was done on synthetic data to have a ground truth to quantify performance of different methods. We have created a wide variety of datasets to cover important aspects of time series data in a classification set up. To show easy-to-interpret saliency maps in addition to the quantified metrics, we used timeseries MNIST [1] over which we demonstrated issues of saliency maps and showed how TSR improves it. To verify that our metrics are reasonable, we calculated saliency distributions on real fMRI time series dataset (supplementary page 3) where we observe similar trends as in synthetic datasets.

Limitations of the new metric & guidance on threshold selection. The limitation of metrics is the need for ground truth, hence the need for synthetic data. Limitation of TSR is cost and depending on the α as an optimization choice that controls the time complexity of TSR (supplementary page 9&10). We will include a more detailed discussion.

Response to Reviewer 3: Problem setup: It is a time series classification setup where all timesteps contribute to making the final output, labels are available after the last timestep (this setup is used in many real-case applications such as an fMRI task classification problem and we are interested in seeing how feature importance change across time[2]). $S(X)$ is network output i.e. $S_1(X)$ is the probability of class 1.

Using gradient-based saliency methods with architectures suffering from vanishing gradients: 3 out of 4 architectures (TCN, Transformer and Input-cell attention proposed by Ismail et al) tested in the benchmark do not suffer from vanishing gradient problem. As discussed, we find that the gradient-based saliency methods limitations for multivariate timeseries is consistent across such architectures in different setups.

Regarding evaluation datasets: Upon your suggestion, we generated a new dataset with the advised setting. Informative feature in new dataset follows a non–linear autoregressive moving average (NARMA) and the non-informative features are Gaussian processes with mean and standard deviation uniformly chosen. The class assignment is generated as the sign of summation of salient features at different informative timesteps. As shown in figure 1, findings are consistent with that of the main paper; the behavior of temporal generated datasets are similar to paper's synthetic datasets. We will include this dataset along with more datasets generated using HMMs and state models in the benchmark.

Use of GradientSHAP for timeseries: GradientShap approximates SHAP values by computing the expectations of gradients by randomly sampling from the distribution of baselines/references. Like others [3,4], we believe such method can be used for timeseries data. We will consider replacing GradientSHAP with SHAP in our final draft.

Response to Reviewer 4: Regarding the proposed approach (TSR). It is an early attempt to tackle the problem observed in all the benchmarked saliency methods which fail to highlight important features within a time step. Our approach shows that TSR improves the quality of the produced saliency maps. However, a more elaborate solution would be essential as soon as a more extensive understanding of the fundamental causes of that problem is achieved. We present this open problem to the community in this paper along with a ready-to-use benchmark with which to approach it.

[1] Bai, et al. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling".
[2] Thomas, Armin W., et al. "Interpretable LSTMs for whole-brain neuroimaging analyses."
[3] Tonekaboni, Sana, et al. "What went wrong and when? Instance-wise Feature Importance for Time-series Models."
[4] Lundberg, Scott, et al."Explainable machine-learning predictions for the prevention of hypoxaemia during surgery.".