Temporal Dependencies in Feature Importance for Time Series Predictions

Clayton Rooke [1 2] **Jonathan Smith** [1] **Kin Kwan Leung** [1] **Maksims Volkovs** [1] **Saba Zuberi** [1]

Abstract

Explanation methods applied to sequential models for multivariate time series prediction are receiving more attention in machine learning literature. While current methods perform well at providing instance-wise explanations, they struggle to efficiently and accurately make attributions over long periods of time and with complex feature interactions. We propose WinIT, a framework for evaluating feature importance in time series prediction settings by quantifying the shift in predictive distribution over multiple instances in a windowed setting. Comprehensive empirical evidence shows our method improves on the previous state-of-the-art, FIT, by capturing temporal dependencies in feature importance. We also demonstrate how the solution improves the appropriate attribution of features within time steps, which existing interpretability methods often fail to do. We compare with baselines on simulated and real-world clinical data. WinIT achieves 2.47× better performance than FIT and other feature importance methods on real-world clinical MIMIC-mortality task. The code for this work is available at https://github.com/layer6ai-labs/WinIT.

1. Introduction

Explaining model predictions is important for transparency, accountability, and for motivating users to act on data. Good methods for generating explanations are particularly useful in domains like healthcare and finance, where explanations are an ethical and legal requirement (Amann et al., 2020). However, the field of time series explainability for deep neural networks has only recently seen attention, with the discovery that traditional explainability methods underperform on deep learning models applied in the time series domain (Ismail et al., 2020). Recent methods such as Feature Importance in Time (FIT) (Tonekaboni et al., 2020) and Temporal Saliency Rescaling (TSR) (Ismail et al., 2020) have improved performance and defined initial benchmarks but face challenges in the breadth of their application in real-world scenarios.

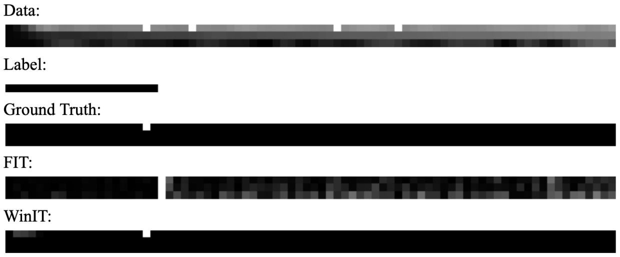


Figure 1. Saliency maps for FIT and WinIT explainability methods on Delayed Spike data (d=2) are compared with the data, per-time step labels and ground truth feature importance. Here it can be seen that WinIT captures the important observation despite a delay between the observation and the label change, unlike FIT, which also overweights all features in the time step with the label change.

In this work we explore time series explainability in the domain where there may be a delay between important feature shifts and a change in the predictive distribution. This type of temporal dependency can be important in real-world settings, where changes in input features may not instantaneously change model predictions. We demonstrate experimentally that existing state-of-the-art method FIT fails to extend to the delayed label setting via experiments on a new synthetic dataset. We propose a new approach, WinIT, to address this challenge by quantifying the impact of features on the predictive distribution over multiple instances in a windowed setting. WinIT utilizes a modification of the instance-wise importance score introduced in FIT, which we refer to as Inverse FIT, that performs better in the windowed setting. We evaluate WinIT on real-world clinical data and find that it outperforms FIT by a significant margin. In summary, our main contributions are:

- Extending FIT to work with lookback-windows that improve performance on datasets where there is some time delay between the observation of important features and a corresponding shift in label. We show how

[1] Layer 6 AI, Toronto, Canada [2] University of Waterloo, Waterloo, Canada. Correspondence to: Clayton Rooke <cjrooke@uwaterloo.ca>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

to evaluate performance on the label delay problem with a new synthetic dataset.

- Reformulating the counterfactual explanation method of FIT in a more efficient manner, suitable for use in a windowed setting.
- Our results show that combining these methods leads to a $2.47\times$ improvement in explanation performance on the real-world clinical MIMIC-mortality task.

2. Background

Traditional perturbation-based and model-based methods have shown limited success in the time series domain. Gradients, Integrated Gradients, GradientSHAP, Deep-LIFT (Shrikumar et al., 2017), and DeepSHAP (Lundberg & Lee, 2017) all leverage model gradients to generate feature importance, but do not directly consider the temporal nature of the problem. Perturbation-based methods like feature occlusion (Zeiler & Fergus, 2014) and feature ablation (Suresh et al., 2017) are model-agnostic methods which measure how changes to the input features relate to changes in model prediction. RETAIN learns attention scores over the input features (Choi et al., 2016). LIME learns explainable models locally around a prediction, applied at every time step in the time series domain (Ribeiro et al., 2016).

Recent benchmarks (Ismail et al., 2020; Tonekaboni et al., 2020) evaluate these traditional explainability methods on time series problems in both simulated and real-world experiments. By separating the importance calculation in both the time and feature input dimensions (Ismail et al., 2020) finds that the performance of the existing methods can be improved. In contrast (Tonekaboni et al., 2020) proposes a new method, FIT, that measures each observation's contribution to the predictive distribution shift of the model over time in order to provide better explanations in certain settings. However, FIT is limited to measuring the importance of instantaneous shifts in the predictive distribution. We define an instantaneous shift as one where the important observations from the input change the model prediction immediately. As in, the important data and the prediction change occur on the same time step. The assumption that feature shift and prediction shift occur simultaneously does not always hold in practice. In real world applications there can be a delay between an important feature shift and a change in outcome. It is important for explanation methods for time series predictions to be able to perform well given such temporal dependencies. For this reason we present a new method, WinIT, which reformulates the FIT algorithm to make it more efficient, while also attributing correct feature importance for non-instantaneous changes in the predictive distribution.

3. Notation

Let $\mathbf{X} \in \mathbb{R}^{D\times T}$ be a sample of a multi-variate time series with D features and T time steps. We denote $[N]$ to be the set $\{1, \ldots, N\}$. We also let $\mathbf{x}_t := \mathbf{X}_{\cdot,t} \in \mathbb{R}^D$ be the set of all observations at a particular time $t \in [T]$ and $\mathbf{X}_{1:t} := [\mathbf{x}_1; \mathbf{x}_2; \ldots ; \mathbf{x}_t] \in \mathbb{R}^{D\times t}$. Let $y_t \in [K]$ be the label at each time step for a classification task with K classes. Let $S \subseteq [D]$ be a subset of features of interest and $\mathbf{x}_{S,t}$ be the observations of that subset at time t. We also define S^c as the set complement of the features of interest. For a model, f_θ, that estimates the conditional distribution $p(y_t|\mathbf{X}_{1:t})$ at each time step, we aim to provide a feature importance score for each set of observations $\mathbf{x}_{S,t}$ using the observations up to that time step, $\mathbf{X}_{1:t} \in \mathbb{R}^{D\times t}$.

For feature importance methods that calculate scores over a set of time steps, we let $n \in [N]$ be the lookback window up to a maximum window size of N. Then $\mathbf{X}_{S,t-n:t}$ represents the set of observations of the subset of features of interest over a set of time steps of length n and $\mathbf{X}_{1:t-n-1}$ represents the historical observations for all features before that window. We also refer to the *absmax* function which in our implementation finds the maximum absolute value, but then returns the actual value, not the absolute value.

4. Methods

In this section we introduce our approach WinIT. We first review the FIT importance score in Section 4.1. We then present Inverse FIT, a modified version of the importance score in Section 4.2. In Section 4.3 we present WinIT, which extends Inverse FIT using a windowed approach to computing feature importance for non-instantaneous changes in the predictive distribution.

4.1. FIT

Proposed by (Tonekaboni et al., 2020), FIT defines an importance score for a subset of features S at time t, given by a set of observations $\mathbf{x}_{S,t}$. It measures how well the partial conditional distribution, where only a subset of features are observed at time t, $p(y|\mathbf{X}_{1:t-1}, \mathbf{x}_{S,t})$, approximates the full predicted distribution $p(y|\mathbf{X}_{1:t})$. This is characterized by the KL divergence between these two distributions and is referred to as the "unexplained" predictive distribution shift. It is measured with respect to the total "temporal" shift from time $t-1$ to time t, given by the KL divergence between the model prediction at time t and the model prediction at time $t-1$, $KL(p(y|\mathbf{X}_{1:t})||p(y|\mathbf{X}_{1:t-1}))$. The FIT importance score for a set of observations S at time t is then given by the difference between the "temporal" distribution shift and the “unexplained” distribution shift:

$$I_{FIT}(\mathbf{x}_S, t) = KL(p(y|\mathbf{X}_{1:t})||p(y|\mathbf{X}_{1:t-1})) - KL(p(y|\mathbf{X}_{1:t})||p(y|\mathbf{X}_{1:t-1}, \mathbf{x}_{S,t})) \quad (1)$$

To compute the partial predictive distribution $p(y|\mathbf{X}_{1:t-1}, \mathbf{x}_{S,t})$, FIT marginalizes over the complement feature set at time t, $\mathbf{x}_{S^c,t}$, by sampling from the counterfactual distribution $p(\mathbf{x}_{S^c,t}|\mathbf{X}_{1:t-1}, \mathbf{x}_{S,t})$ approximated by a generative model G.

4.2. Inverse FIT

The FIT algorithm quantifies the predictive distribution shift explained by an observation $\mathbf{x}_{S,t}$ by calculating the difference between the unexplained distribution shift and the total temporal distribution shift. An alternative approach is to directly compute the explained distribution shift. We can measure the importance of features S at time t by quantifying how well the partial conditional distribution, $p(y|\mathbf{X}_{1:t-1}, \mathbf{x}_{S^c,t})$, where only the complement set of features, S^c, are observed at time t, approximates the true predictive distribution $p(y|\mathbf{X}_{1:t})$. We call this modification Inverse FIT (IFIT). The new formulation of an instance-wise feature importance score is:

$$I_{IFIT}(\mathbf{x}_S, t) = KL(p(y|\mathbf{X}_{1:t})||p(y|\mathbf{X}_{1:t-1}, \mathbf{x}_{S^c,t})) \quad (2)$$

Similar to FIT, we compute the partial predictive distribution $p(y|\mathbf{X}_{1:t-1}, \mathbf{x}_{S^c,t})$ by using Monte-Carlo integration to marginalize over $\mathbf{x}_{S,t}$ by sampling from a generator G that approximates the distribution $p(\mathbf{x}_{S,t}|\mathbf{X}_{1:t-1}, \mathbf{x}_{S^c,t})$. This approach is outlined in Algorithm 1.

It is important to note that the Inverse Fit importance score, Equation 2 is not equivalent to the FIT score. In particular Inverse FIT does not consider the overall shift in the predictive distribution from $t-1$ to t. This approach performs well when extended to calculating feature importance over a window of time steps, as shown in Section 5. Furthermore, we find that Inverse FIT achieves similar performance to FIT, but is faster as seen in Table 1. This is due to the different generator that can be used (per-feature, rather than joint) and is relevant in the case where $|S| = 1$ as evaluated in our experiments. For larger set sizes the runtime may vary.

FIT is limited to measuring instantaneous changes in the predictive distribution, because only the most recent time step of input is considered when computing importance for a given prediction. The importance score $I_{FIT}(\mathbf{x}_S, t)$ in Equation 1 is equal to the predictive distribution shift from time step $t-1$ to t explained by the observation $\mathbf{x}_{S,t}$. However, for sequential models, the observation $\mathbf{x}_{S,t}$ could also influence any of the predictive distributions from time $t+1$ onwards. This importance is not captured in the FIT algorithm. In the next section we extend the IFIT method to address this limitation.

Algorithm 1 IFIT

Input: f_θ: Trained Black-box predictor model, G: Trained generative model, $\mathbf{X}_{1:T} \in \mathbb{R}^{D \times T}$: Time series where T is the max time and D is the number of features, S: a subset of features of interest, L: number of Monte Carlo samples

Output: Importance score matrix $I \in \mathbb{R}^{T \times D}$

```
Train $G$ using $\mathbf{X}_{1:T}$
for all $t \in [T]$ do
    $p(y|\mathbf{X}_{1:t}) = f_\theta(\mathbf{X}_{1:t})$
    $p(\mathbf{x}_t|\mathbf{X}_{1:t-1}) \approx G(\mathbf{X}_{1:t-1})$
    for all $l \in [L]$ do
        Sample $\hat{\mathbf{x}}_{S,t}^{(l)} \sim p(\mathbf{x}_{S,t}|\mathbf{X}_{1:t-1}, \mathbf{x}_{S^c,t})$
        $p(\hat{y}^{(l)}) = f_\theta(\mathbf{X}_{1:t-1}, \mathbf{x}_{S^c,t}, \hat{\mathbf{x}}_{S,t}^{(l)})$
    end for
    $p(y|\mathbf{X}_{1:t-1}, \mathbf{x}_{S^c,t}) \approx \frac{1}{L}\sum_{l=1}^{L} p(\hat{y}^{(l)})$
    $I(\mathbf{x}_S, t) = KL(p(y|\mathbf{X}_{1:t})||p(y|\mathbf{X}_{1:t-1}, \mathbf{x}_{S^c,t}))$
end for
```

4.3. WinIT

We formulate an extension of IFIT with a window of past observations when attributing importance for a given prediction and call this WinIT. For a prediction at time t, with a window size of N, we compute importance scores for the observations $\mathbf{X}_{S,t-N:t}$. For a set of observations $\mathbf{x}_{S,t-n}$ the sum of the I_{IFIT} importance scores for all remaining time steps $t-n+1$ to t is subtracted from the total importance score for time steps $t-n$ to t, to get the observation score at time $t-n$. When $n = 1$ the importance score for remaining time steps is zero. Because the KL divergences in a sequence of n windows cancel out in subsequent scores this can be rewritten as the difference between the current time step and the following time step as seen in in Equation 3. Here W represents the importance of the feature subset S at time step $t-n$ that affects the prediction at time step t.

Since WinIT generates N scores for each time step, but the explainability methods were evaluated based on single importance score for each observation on individual features (no subsets), the final scores must be aggregated. To generate a single observation score, the absolute maximum value across all of the $n \in [N]$ windows is computed as shown in Equation 4. This is used in order to capture the most important contribution of each observation in the final importance score. In the benchmark experiments, important contributions tend to be sparse; a different aggregation metric like an average would not properly capture infrequent but important contributions.

Algorithm 2 WinIT

Input: f_θ: Trained Black-box predictor model, G: Trained generative model, $\mathbf{X}_{1:T} \in \mathbb{R}^{D\times T}$: Time series where T is the max time and D is the number of features, N: feature importance lookback window size, S: a subset of features of interest, L: number of Monte Carlo samples

Output: Importance score matrix $I \in \mathbb{R}^{T\times D\times N}$

```
Train $G$ using $\mathbf{X}_{1:T}$
for all $t \in [T]$ do
  $p(y|\mathbf{X}_{1:t}) = f_\theta(\mathbf{X}_{1:t})$
  $KL_0 = 0$
  for all $n \in [N]$ do
    $p(\mathbf{x}_{t-n:t}|\mathbf{X}_{1:t-n-1}) \approx G(\mathbf{X}_{1:t-n-1})$
    for all $l \in [L]$ do
      Sample $\hat{\mathbf{X}}^{(l)}_{S,t-n:t} \sim p(\mathbf{X}_{S,t-n:t}|\mathbf{X}_{1:t-n-1}, \mathbf{X}_{S^c,t-n:t})$
      $p(\hat{y}^{(l)}) = f_\theta(\mathbf{X}_{1:t-n-1}, \mathbf{X}_{S^c,t-n:t}, \hat{\mathbf{X}}^{(l)}_{S,t-n:t})$
    end for
    $p(y|\mathbf{X}_{1:t-n-1}, \mathbf{X}_{S^c,t-n:t}) \approx \frac{1}{L}\sum_{l=1}^{L} p(\hat{y}^{(l)})$
    $KL_n = KL(p(y|\mathbf{X}_{1:t})||p(y|\mathbf{X}_{1:t-n-1}, \mathbf{X}_{S^c,t-n:t}))$
    $W(\mathbf{x}_S, t-n, n) = KL_n - KL_{n-1}$
    $KL_{n-1} = KL_n$
  end for
end for
$I_{WinIT}(\mathbf{x}_S, t) = \text{absmax}_{n\in[N]} W(\mathbf{x}_S, t, n)$
Return $I$
```

$$W(\mathbf{x}_S, t, n) = KL(p(y|\mathbf{X}_{1:t})||p(y|\mathbf{X}_{1:t-n-1}, \mathbf{X}_{S^c,t-n:t})) - KL(p(y|\mathbf{X}_{1:t})||p(y|\mathbf{X}_{1:t-n}, \mathbf{X}_{S^c,t-n+1:t})) \tag{3}$$

$$I_{WinIT}(\mathbf{x}_S, t) = \text{absmax}_{n\in[N]} W(\mathbf{x}_S, t, n) \tag{4}$$

This leads to a new formulation of a instance-wise feature importance score now taken over multiple overlapping windows described in Algorithm 2.

5. Experiments

For the following experiments, an RNN-based predictor is trained on the training dataset, and the explainability methods are evaluated on the test dataset. A recurrent latent variable generator (Chung et al., 2015) is trained on the training dataset for the FIT and WinIT models.

To evaluate the explainability methods on experiments with simulated data ground truth importance scores are defined. An observation is given a ground truth importance score of 1 if it causes the label to change. All other observations have a ground truth importance score of 0. Explainability methods are evaluated against the ground truth using AUROC and AUPRC, where the ranking score is calculated per-sample by ranking all the feature instances by importance and then averaged over the entire dataset. For the real-world clinical data, no ground truth feature importance is available. Instead, the explainability methods are evaluated based on AUROC drop after the Top K=50, or Top 5%, of observations with the highest importance score is removed from the test dataset by carrying forward the previous values.

Table 1. Explanation performance on Spike, and Delayed Spike (d=2) datasets. For WinIT we use a window size of 8. All evaluations are conducted over 5 random seeds and averaged.

SPIKE

METHOD	AUROC	AUPRC	TIME (S)
FIT	0.994 ± 0.002	0.852 ± 0.098	394.78
IFIT	0.954 ± 0.006	0.844 ± 0.081	70.91
WINIT	0.965 ± 0.002	0.905 ± 0.048	449.41

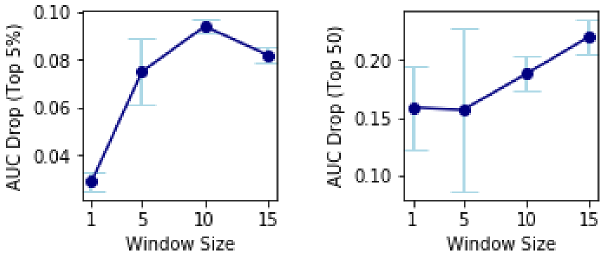
DELAYED SPIKE (D=2)

METHOD	AUROC	AUPRC	TIME (S)
FIT	0.516 ± 0.035	0.002 ± 0.001	340.00
WINIT	0.970 ± 0.006	0.909 ± 0.029	455.03

5.1. Simulated Data

Spike is a benchmark experiment presented in (Tonekaboni et al., 2020) which uses a multivariate dataset composed of 3 random NARMA time series with random 'spikes', immediate large increases, added to the samples. The label is 0 until a spike occurs in the first feature, at which point it changes to 1 for the rest of the sample. As shown in Table 1, WinIT shows similar performance to FIT on the Spike benchmark, with FIT having the highest AUROC, and the AUPRC of the two methods being the same within one standard deviation.

To demonstrate the temporal dependency effect, we present a simple experiment using simulated data as a modification of the Spike data. Three independent NARMA sequences are generated and two of the features add linear trends. Spikes are then added following the same procedure as the Spike data. However, the time step at which the label changes is different. In the Spike data the label changes immediately to 1 after encountering the first spike. In the Delayed Spike data the label changes after $d = 2$ time steps. To measure the accuracy of the explainability methods, we define the ground truth importance score as 1 for the first spike and 0 for all other observations. As shown in Table 1, since FIT only considers the observations from time steps $t-1$ to t as they relate to the prediction at t, it is unable to assign importance to the correct observation, which occurs



(a) (b)

Figure 2. WinIT performance on the MIMIC-mortality task improves as the window size increases from 1 to 15 for AUC Drop (Top 5%) (a) and AUC Drop (K=50) (b).

at time step $t - d$. However, since the spike falls in the window $\mathbf{X}_{t-N:t}$, WinIT is able to assign importance to the correct observation.

We also show performance of the IFIT and WinIT models as an ablation study in Table 1. This reveals some of the tradeoffs between runtime and performance for the different methods, with IFIT-based methods taking less time than FIT. Methods that do not use a lookback window achieve poor results on the Delayed Spike data, reflected in the low AUPRC of both IFIT and FIT.

5.2. Clinical Data

MIMIC III is a multivariate time series clinical dataset with a number of vital and lab measurements taken over time for around 40000 patients at the Beth Israel Deaconess Medical Center in Boston, MA (Johnson et al., 2016). MIMIC III is used in the FIT paper to construct the MIMIC-mortality experiment, which uses 8 vital and 20 lab measurements hourly over a 48 hour period to predict patient mortality. As shown in Table 2, the WinIT method is a significant improvement over FIT and other explainability methods on the MIMIC-mortality experiment. In fact, we see a $2.47\times$ improvement over FIT when calculating AUC Drop in the top 5% of features and a $1.36\times$ improvement when calculating AUC Drop in the top 50 features.

Adjusting the window size can lead to different performance in all settings. WinIT with different lookback windows of size 1, 5, 10, and 15 shows improving performance in AUC Drop (Top 5%) in the real-world setting of the MIMIC-mortality task as seen in Figure 2. AUC Drop (Top 50), while outperforming all other methods, does exhibit more variance and does not improve with window size. This may be because only a few features benefit from the additional information related to delays between feature changes and label changes. It may also be that globally important features from the Top 5% display more temporal dependence than other features.

Table 2. Explanation performance on MIMIC-mortality task. WinIT uses a window size of 10. At the bottom we show performance improvement against the second-best method (FIT). † indicates results are from (Tonekaboni et al., 2020).

METHOD	AUC DROP (95-PC)	AUC DROP (K=50)
AFO†	0.023 ± 0.003	0.068 ± 0.003
FO†	0.028 ± 0.006	0.095 ± 0.042
DEEP-LIFT†	0.045 ± 0.004	0.067 ± 0.038
IG†	0.036 ± 0.003	0.056 ± 0.014
RETAIN†	0.020 ± 0.014	0.032 ± 0.019
LIME†	0.028 ± 0.000	0.032 ± 0.019
GRADSHAP†	0.036 ± 0.000	0.065 ± 0.062
FIT	0.038 ± 0.005	0.138 ± 0.037
WINIT	$\mathbf{0.094 \pm 0.003}$	$\mathbf{0.188 \pm 0.015}$
(VS. FIT)	$2.47\times$	$1.36\times$

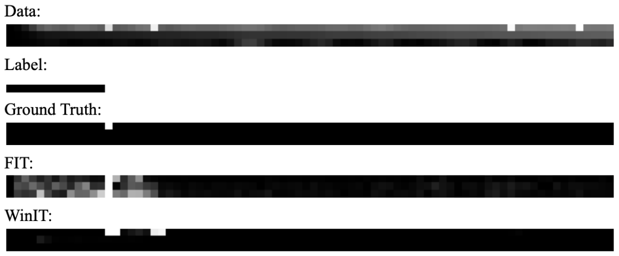


Figure 3. Saliency maps for FIT and WinIT methods on the Spike dataset. Here it can be seen that both methods capture the important observation, but FIT also overweights the importance of all features in the important time step and other time steps for all features, while WinIT only overweights a few other time steps in the important feature.

5.3. Saliency Maps

As a sanity check on the explanations provided by WinIT we show saliency maps from instances in the Spike and Delayed Spike datasets in Figure 1 and Figure 3. In the Delayed Spike example it is clear how FIT fails to identify the important observations, instead providing a higher average score to all observations. FIT also suffers from a common problem in time series explanations, where it overweights features that occur in the same time step as an important observation. This can be seen in both figures. WinIT, on the other hand, sees failure cases for the Spike dataset when multiple spikes appear close together.

We also show comparisons of FIT and WinIT saliency maps for an instance from the MIMIC-mortality task in Figure 4. In this case the overweighting of time steps is even more apparent with the FIT explanation due to the larger number of features.

6. Conclusion

In this work we propose WinIT, a method for time series explainability that allows for attributing correct importance to observations for non-instantaneous changes in a time series model's predictive distribution. WinIT uses a windowed approach to computing the feature importance and is based on a modification of the FIT importance score that performs well in the windowed setting. WinIT is comparable to FIT on the Spike benchmark, and significantly outperforms FIT on the proposed Delayed Spike data,where changes in the model's predictive distribution are not instantaneous, as well as on the real-world MIMIC-mortality task.

In the future, we hope to evaluate WinIT on the other benchmark experiments, as well as new simulated and real-world experiments to help better understand where temporal dependencies make the greatest impact. The methods we present can be further optimized through selection and tuning of the generative methods used and their application to different kinds of real-world data.

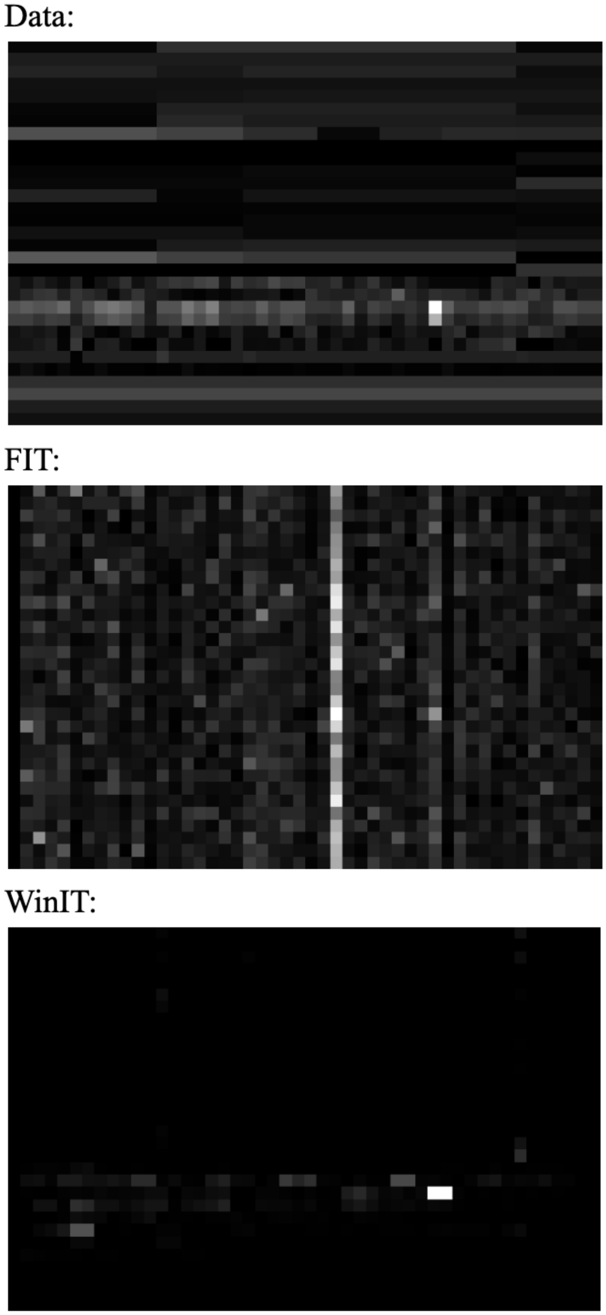


Figure 4. Saliency maps for FIT and WinIT methods on an instance from the MIMIC-mortality task. Here the features identified as most important by WinIT have stronger correlation with the underlying features while FIT shows an over weighting of a single time step.

References

Amann, J., Blasimme, A., Vayena, E., Frey, D., and Madai, V. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(310), 2020.

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/231141b34c82aa95e48810a9d1b33a79-Paper.pdf.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., and Bengio, Y. A recurrent latent variable model for sequential data. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/b618c3210e934362ac261db280128c22-Paper.pdf.

Ismail, A. A., Gunady, M., Corrada Bravo, H., and Feizi, S. Benchmarking deep learning interpretability in time series predictions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6441–6452, 2020. URL https://proceedings.neurips.cc/paper/2020/file/47a3893cc405396a5c30d91320572d6d-Paper.pdf.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, 2017.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. *CoRR*, abs/1704.02685, 2017. URL http://arxiv.org/abs/1704.02685.

Suresh, H., Hunt, N., Johnson, A. E. W., Celi, L. A., Szolovits, P., and Ghassemi, M. Clinical intervention prediction and understanding using deep networks. *CoRR*, abs/1705.08498, 2017. URL http://arxiv.org/abs/1705.08498.

Tonekaboni, S., Joshi, S., Campbell, K., Duvenaud, D. K., and Goldenberg, A. What went wrong and when? instance-wise feature importance for time-series black-box models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 799–809, 2020. URL https://proceedings.neurips.cc/paper/2020/file/08fa43588c2571ade19bc0fa5936e028-Paper.pdf.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. *European conference on computer vision*, 2014.