# Temporal Dependencies and Spatio-Temporal Patterns of Time Series Models

**Md. Khairul Islam**

Computer Science, University of Virginia
Charlottesville, Virginia, USA.
mi3se@virginia.edu

## Abstract

The widespread use of Artificial Intelligence (AI) has highlighted the importance of understanding AI model behavior. This understanding is crucial for practical decision-making, assessing model reliability, and ensuring trustworthiness. Interpreting time series forecasting models faces unique challenges compared to image and text data. These challenges arise from the temporal dependencies between time steps and the evolving importance of input features over time. My thesis focuses on addressing these challenges by aiming for more precise explanations of feature interactions, uncovering spatiotemporal patterns, and demonstrating the practical applicability of these interpretability techniques using real-world datasets and state-of-the-art deep learning models.

## Introduction

The reliability of AI models is essential for their widespread use. Explanations provide the transparency and aid needed to make reliable decisions. Especially in sensitive data, it is often an ethical and legal requirement. Higher interpretability leads to several key benefits: 1) More acceptance and trust in the system's decisions 2) Revealing incompleteness in the problem formalization and debugging 3) Improved scientific understanding of the problem 4) Reliable and better performance.

The large number of real-world applications of deep learning time series forecasting models in areas like medicine, finance, retail, and traffic, better interpretability of these models has significant practical implications (Rojat et al. 2021; Benidis et al. 2022). Explaining time series models with interpretation methods can highlight the importance of input features to the model's prediction. Different time series tasks doing classification, regression, anomaly detection, or missing value imputation require carefully designed techniques to explain the model behavior. Both black-box and white-box interpretation techniques can be developed for this.

My work on time series interpretation is organized into three key contribution phases. In Phase I (Section ), I delve into interpreting patterns that the model learns using its built-in structure, like examining temporal self-attention weights. Phase II (Section ) overcomes the model-specific limitation of Phase I by designing black-box interpretation methods and focusing on the temporal feature importance. Phase III (Section ) aims to further improve and generalize the black-box interpretation methods for time series models given past observations over a fixed time window. As used by the latest Transformer-based models (Benidis et al. 2022).



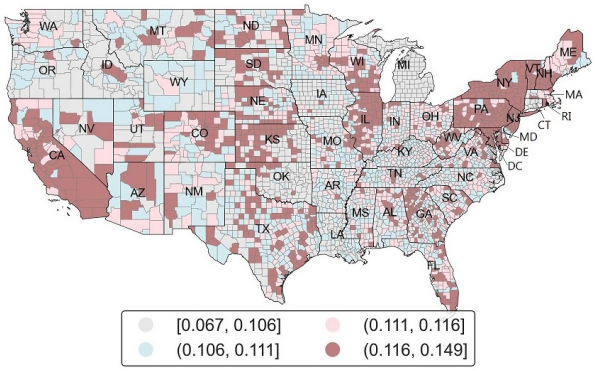


Figure 1: Average attention weights from TFT for all US counties (Islam et al. 2023a).

## Phase I: Interpreting Spatio-Temporal Patterns with Self-Attention Weights

This phase investigates the research question: What kind of spatio-temporal patterns are learned by a deep learning model and how to explain them? I interpreted daily COVID-19 infection predictions for 3,142 US counties (Islam et al. 2023a) with 2.5 years of collected data. We used the previous 13 days of input to predict the next 15 days of COVID-19 cases for each county. The evaluation using five regression metrics showed that the Temporal Fusion Transformer (TFT) (Lim et al. 2021) outperformed the other four deep learning models in all metrics.

I used the self-attention weights extracted from the TFT model to show the different spatio-temporal patterns learned by the model. I found they are correlated to the raw data patterns and are caused by the way COVID-19 cases were reported by the health sectors and the pandemic dynamics. Figure 1 shows the aggregated attention map over our study

period. The spatial interpretation can highlight the points of significance with county-level granularity and the clusters of counties with more impacted by the epidemic.

## Phase II: Interpreting Feature Interactions Using Sensitivity Analysis

In this phase, I aim to explain the direct interactions between input features and output targets, using black-box interpretation methods. This overcomes the limitation of the previous phase, where the self-attention weights only catch the higher-level patterns from the encoded representation of the dataset.

I investigated using different sensitivity analysis methods to quantify the contribution of different age groups to COVID-19 infection spread (Islam et al. 2023b). This is important for the administration to investigate which age groups contribute more to the infection spread. We collected 2 years of COVID-19 infection data at the US county level.

The age groups are defined based on the Centers for Disease Control and Prevention's (CDC) report on COVID-19 cases for different age groups. I trained the TFT (Lim et al. 2021) model on the dataset and calculated the sensitivity of the age group features with respect to the model. The sensitivity ranks compared with the age group cases reported by the CDC showed high accuracy of our interpretation method.

## Phase III: Windowed Temporal Saliency Rescaling

In this step, I focus on how well we can interpret the temporal feature importance of time series models and their temporal dependencies with a fixed input context. Calculating the change of feature relevance with time is the key to understanding the contribution of those features, their interactions, choosing important features, and finding shifting points (Leung et al. 2022; Ozyegen, Ilic, and Cevik 2022).

Recent benchmarks have shown that interpretation methods underperform in the time domain (Ismail et al. 2020) for not considering the temporal order and relation of the data. Techniques that aim to highlight important observations by treating them independently face significant limitations when the same observation can vary in importance to predictions at different times, defined as a temporal dependency.

Another limitation of some prior works is that they use the last time step to calculate the input feature importance (Ismail et al. 2020; Tonekaboni et al. 2020). But when there is a delay between the important feature shifts and a change in the target output, those dynamics are difficult to capture using such an approach (Leung et al. 2022).

To address these challenges, I propose a novel windowed temporal saliency rescaling technique that determines the importance of a given observation in time to a prediction horizon over the input look-back window. Compared to the related works (Ozyegen, Ilic, and Cevik 2022; Leung et al. 2022), this solution calculates the importance of each input time step separately, then rescales the feature importance across different lookback positions using it. Hence doesn't assume the sole importance of the last time step and considers temporal dependency. In short, I plan to,

- Collect multiple datasets covering both classification and regression tasks. Train deep learning models with different architectures on these datasets in a multi-horizon setting with a fixed look-back window.
- Use the proposed novel algorithm to calculate the importance of the input features over the input window. Benchmark the performance using recent local interpretation methods for time series.

| Phase | Completed? | Est. Remaining |
|---|---|---|
| Phase-I | Yes | - |
| Phase-II | No | 10% |
| Phase-III | No | 20 % |

Table 1: Progress Summary as of December 1, 2023.

## References


Benidis, K.; Rangapuram, S. S.; Flunkert, V.; Wang, Y.; Maddix, D.; Turkmen, C.; Gasthaus, J.; Bohlke-Schneider, M.; Salinas, D.; Stella, L.; et al. 2022. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55(6): 1–36.

Islam, M. K.; Liu, Y.; Erkelens, A.; Daniello, N.; Marathe, A.; and Fox, J. 2023a. Interpreting County-Level COVID-19 Infections using Transformer and Deep Learning Time Series Models. In *2023 IEEE International Conference on Digital Health (ICDH)*, 266–277.

Islam, M. K.; Valentine, T.; Wang, R.; Davis, L.; Manner, M.; and Fox, J. 2023b. Population Age Group Sensitivity for COVID-19 Infections with Deep Learning. *arXiv preprint arXiv:2307.00751*.

Ismail, A. A.; Gunady, M.; Corrada Bravo, H.; and Feizi, S. 2020. Benchmarking deep learning interpretability in time series predictions. *Advances in neural information processing systems*, 33: 6441–6452.

Leung, K. K.; Rooke, C.; Smith, J.; Zuberi, S.; and Volkovs, M. 2022. Temporal Dependencies in Feature Importance for Time Series Prediction. In *The Eleventh International Conference on Learning Representations*.

Lim, B.; Arık, S. Ö.; Loeff, N.; and Pfister, T. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764.

Ozyegen, O.; Ilic, I.; and Cevik, M. 2022. Evaluation of interpretability methods for multivariate time series forecasting. *Applied Intelligence*, 1–17.

Rojat, T.; Puget, R.; Filliat, D.; Del Ser, J.; Gelin, R.; and Díaz-Rodríguez, N. 2021. Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950*.

Tonekaboni, S.; Joshi, S.; Campbell, K.; Duvenaud, D. K.; and Goldenberg, A. 2020. What went wrong and when? Instance-wise feature importance for time-series black-box models. In *Neural Information Processing Systems*.