Interpreting Time Series Transformer Models and Sensitivity Analysis of Population Age Groups to COVID-19 Infections

Md Khairul Islam [1], Tyler Valentine [2], Timothy Joowon Sue [1], Ayush Karmacharya [1], Luke Neil Benham [1], Zhengguang Wang [1], Kingsley Kim [1], Judy Fox [1, 2]

[1] Computer Science, University of Virginia, USA.
[2] School of Data Science, University of Virginia, USA.
Email: {mi3se, xje4cy, und2yw, psb7wm, lnb6grp, zw4re, bjb3az, cwk9mp}@virginia.edu

Abstract

Interpreting deep learning time series models is crucial in understanding the model's behavior and learning patterns from raw data for real-time decision-making. However, the complexity inherent in transformer-based time series models poses challenges in explaining the impact of individual features on predictions. In this study, we leverage recent local interpretation methods to interpret state-of-the-art time series models. To use real-world datasets, we collected three years of daily case data for 3,142 US counties. Firstly, we compare six transformer-based models and choose the best prediction model for COVID-19 infection. Using 13 input features from the last two weeks, we can predict the cases for the next two weeks. Secondly, we present an innovative way to evaluate the prediction sensitivity to 8 population age groups over highly dynamic multivariate infection data. Thirdly, we compare our proposed perturbation-based interpretation method with related work, including a total of eight local interpretation methods. Finally, we apply our framework to traffic and electricity datasets, demonstrating that our approach is generic and can be applied to other time-series domains.

1 Introduction

As the promise of deep learning models in critical domains (Zhao et al. 2023) such as health, finance, and social science, ensuring the interpretability of these methods becomes essential for maintaining AI transparency and the reliability of model decisions (Amann et al. 2020).

The interpretability methods evaluate different factors contributing to the model decisions (Rojat et al. 2021), reveal incompleteness in the problem formalization, and improve our scientific understanding (Doshi-Velez and Kim 2017). Explaining time series models in a meaningful way is challenging due to their dynamic nature. Much research on time series models has focused on interpreting classification tasks or using simple models. It is desirable to understand how to use these interpretation methods in state-of-the-art time series models while still achieving the best performance.

In summary, the main contributions of our research are:

- Use the highly dynamic data in a multivariate, multi-horizon, and multi-modal setting with state-of-the-art time series transformer models. COVID-19 is a recent pandemic taking millions of lives and causing many research efforts to forecast the infection spread using statistical learning, epidemiological, and machine learning models (Clement et al. 2021).
- Collect around three years of COVID-19 data daily for 3,142 US counties. Each county contributes to one time series in the dataset (hence multi-time series). We use the last 14 days of data to predict the cases for the next 14 days. The best-performing model on the test set is later used for interpretation. This allows us to give a more granular analysis.
- Focus on local interpretation methods to show the contribution of each input feature to the prediction given an input sample, and benchmark our interpretation using eight recent methods. Then evaluate the interpreted attribution scores following the latest practices (Ozyegen, Ilic, and Cevik 2022).
- Propose an innovative way to evaluate sensitivities of age group features using infection by age group. We employ a black-box interpretation method in our approach, which applies to various models and time series datasets.

The rest of the sections are organized as follows: Section 2 describes the background and related work. Section 3 defines the problem statement of both forecasting and interpretation tasks. Section 4 describes the data collection and pre-processing steps. Section 5 summarizes the experimental setup, training steps, and interpretation without ground truth for all of our three datasets. Section 6 discusses evaluating age group sensitivity with ground truth for our COVID dataset. Sections 7 and 8 examine additional aspects of our approach and conclude our work.

2 Background and Related Work

In this study, we focus on local interpretation methods consisting of both time series interpretation and the prediction of COVID-19 infection. The following sections contain the related terminologies and recent works on this topic.

2.1 Background Terminologies

Interpretation methods are either: 1)**White box**: Using the model's inherent architecture to interpret model behavior (e.g. using attention weights). 2) **Black box**: Using only

the input and output to determine the model's behavior. We use black box methods in this work since they are model-agnostic and more applicable.

Based on application scope, interpretability methods are either: 1) **Global**: Explains the entire behavior of the model 2) **Local**: Explains the reasons behind a specific model decision on an input instance (Rojat et al. 2021). We focus on local interpretation methods in this work, to provide a granular analysis of the model's behavior.

Interpretation methods aim to quantify the relevance of input features to model output. This helps identify key features influencing the model's decision. Evaluating these importance scores in practice is difficult due to the lack of ground truth for interpretation. However, existing studies (Rojat et al. 2021; Ismail et al. 2020) perturb the top features based on these interpreted importance scores and recalculate how much that impacts the model's output to evaluate interpretation quantitatively (Section 3.1).

Sensitivity analysis is one type of perturbation-based technique to interpret a model's behavior. Morris method (Morris 1991) is a sensitivity analysis method that defines the sensitivity of a model input as the ratio of the change in an output variable to the change in an input feature. We expanded the Morris method to temporal data using the SALib library (Iwanaga, Usher, and Herman 2022). We use the 'mu_star' as the feature importance score, as it is more reliable. A higher 'mu_star' indicates higher sensitivity.

2.2 Related Work

In this study, we focus on local interpretation methods consisting of both time series interpretation and the prediction of COVID-19 infection.

Time Series Interpretation A wide range of interpretation methods has been proposed in the literature (Rojat et al. 2021; Turbé et al. 2023). Including *gradient based methods* such as Integrated Gradients (Sundararajan, Taly, and Yan 2017), GradientSHAP (Erion et al. 2019) which uses the gradient of the model predictions to input features to generate importance scores. *Feature removal based* methods such as Feature Occlusion (Zeiler and Fergus 2013), Feature Ablation (Suresh et al. 2017), and Sensitivity Analysis (Morris 1991) replace a feature or a set of features from the input using some fixed baselines or generated samples and measure the importance based on the model output change.

These methods have been popular and used in time series datasets (Ozyegen, Ilic, and Cevik 2022; Zhao et al. 2023; Turbé et al. 2023). Ismail et al. (2020) standardized the evaluation of local interpretations when no interpretation ground truth is present. Most prior works on time series interpretation focus on classification (Turbé et al. 2023; Ismail et al. 2020; Rojat et al. 2021). Ozyegen, Ilic, and Cevik (2022) proposed a novel evaluation metric for local interpretation in time series regression tasks using real-world datasets.

One key limitation of the prior works is using baseline models or synthetic datasets, which doesn't reflect the practical case where we want to use state-of-the-art models with complex real-world datasets. We address this limitation by incorporating recent time series models in our work.

Interpreting COVID-19 Infection DeepCOVID (Rodriguez et al. 2021) utilized RNN with auto-regressive inputs to predict COVID-19 cases. Then recursively eliminating input signals to quantify the model output deviation without those signals and use that to rank the signal importance. DeepCOVIDNet (Ramchandani, Fan, and Mostafavi 2020) classified infected regions with high, medium, and low case growth, then interpreted using Feature Occlusion on part of the training data.

COVID-EENet (Kim et al. 2022) interpreted the economic impact of COVID-19 on local businesses. Self-Adaptive Forecasting (Arik, Yoder, and Pfister 2022) used the model's attention weights to interpret state-level death forecasts. However, this is model-dependent and can't be applied to models without the attention mechanism.

One key limitation of these works is not comparing their interpretation performance with other interpretation methods. In this work, we address this challenge and bridge the gap by comparing eight recent local interpretation methods.

3 Problem Statement

We consider a multivariate multi-horizon time series setting with length T, the number of input features J, and total N instances. $X_{j,t} \in \mathbb{R}^{J \times T}$ is the input feature j at time $t \in \{0, \cdots, T-1\}$. We use past information within a fixed look-back window L, to forecast for the next τ_{max} time steps. The target output at time t is y_t. Hence our black-box model f can be defined as $\hat{y}_t = f(X_t)$ where,

$$\begin{aligned} X_t &= x_{t-(L-1):t} \\ &= [x_{t-(L-1)}, x_{t-(L-2)}, \cdots, x_t] \\ &= \{x_{j,l,t}\},\ j \in \{1, \cdots, J\},\ l \in \{1, \cdots, L\} \end{aligned} \tag{1}$$

$\hat{y}_t$ is the forecast at $\tau \in \{1, \cdots, \tau_{max}\}$ time steps in the future. X_t is the input slice at time t of length L. An individual covariate at position (n, l) in the full covariate matrix at time step t is denoted as $x_{j,l,t}$.

For interpretation, our target is to construct the importance matrix $\phi_t = \{\phi_{j,l,t}\}$ for each output $o \in O$ and prediction horizon $\tau \in \{1, \cdots, \tau_{max}\}$. So this is a matrix of size $O \times \tau_{max} \times J \times L$. We find the relevance of the feature $x_{j,l,t}$ by masking it in the input matrix X_t and output change from the model,

$$\phi_{j,l,t} = |(f(X_t) - f(X_t \setminus x_{j,l,t})| \tag{2}$$

where $X_t \setminus x_{j,l,t}$ is the feature matrix achieved after masking entry $x_{j,l,t}$.

3.1 Methodology for Local Interpretation of Time-Series

A major challenge in evaluating interpretation is the lack of interpretation of ground truth. We use the following quantitative analysis steps (Ozyegen, Ilic, and Cevik 2022) to **perform *local interpretation evaluation* in the absence of ground truth**:

1. Sort relevance scores $R(X)$ returned by the interpretation method so that $R_e(X_{i,t})$ is the e^{th} element in the ordered set $\{R_e(x_{i,t})_{e=1}^{L \times N}\}$. Here L is the look-back window and N is the number of features.

2. Find top $k\%$ (we used, $k \in \{5, 7.5, 10, 15\}$) entries in this set, where $\mathbf{R}(x_{i,t}) \in \{\mathbf{R}_e(x_{i,t})\}_{e=1}^{k}$.
3. Mask these top features or every other feature.
4. Calculate the change in the model's output to the original output using the mean absolute error (MAE) metric following (Ozyegen, Ilic, and Cevik 2022).

DeYoung et al. (2019) propose the *comprehensiveness* and *sufficiency* metrics to measure the faithfulness of the interpretations. *Comprehensiveness* defines whether all features needed to make a prediction were selected, measured by calculating the output change after masking the top important features. Intuitively, the model should be less confident in its prediction afterward. *Sufficiency* defines whether the features selected as important contain enough information to make the prediction. This is measured by masking any other feature except the top important features. The smaller the output change, the more sufficient the selected features are.

In summary, **the higher the comprehensiveness loss and the lower the sufficiency loss the better**. We define the set of top $k\%$ relevant features selected by the interpretation method for the i-th input X_i as $X_{i,1:k}$, input after removing those features $X_{i,\setminus 1:k}$. Then for our model $f()$ we can describe comprehensiveness and sufficiency as:

$$\begin{aligned} \text{Comprehensiveness} &= |f(X_i) - f(X_{i,\setminus 1:k})| \\ \text{Sufficiency} &= |f(X_i) - f(X_{i,1:k})| \end{aligned} \tag{3}$$

For K bins of top $k\%$ features (we use top 5%, and 10% features, hence $K = 2$.), the aggregated comprehensiveness score is referred to as the "Area Over the Perturbation Curve for Regression" or AOPCR (Ozyegen, Ilic, and Cevik 2022).

$$AOPCR = \frac{1}{K \times \tau_{max}} \sum_{\tau}^{\tau_{max}} \sum_{k}^{K} |f(X_i)_\tau - f(X_{i,\setminus 1:k})_\tau)| \tag{4}$$

We calculate the AOPCR for sufficiency similarly after replacing $X_{i,\setminus 1:k}$ with $X_{i,1:k}$. Table 6 presents the AOPCR results from the interpretation methods. It shows our implementation of the Morris Sensitivity method performing the best in most cases (3 out of 4).

3.2 Interpretation Methods

In this section, we describe our interpretation methods and how to evaluate the interpretation performance across different methods. The interpretation is done using the FEDformer model on the test set. However, the approach is generic and model-agnostic. Therefore it can be used for other time series models.

We use the following recent methods to perform black-box local interpretation analysis on the target mode:

1. ***Feature Ablation (FA):*** Computes (Suresh et al. 2017) attribution as the difference in output after replacing each feature with a baseline.
2. ***Feature Permutation (FP):*** Permutes the (Molnar 2020) the input feature values within a batch, and computes the difference between original and shuffled outputs.
3. ***Morris Sensitivity (MS):*** Morris method (Morris 1991) calculates the model output change with respect to a δ change to the input value. We designed a temporal adaptation of this Morris method using the Sensitivity Analysis Library (Iwanaga, Usher, and Herman 2022).
4. ***Feature Occlusion (FO):*** Replaces the input features with a counterfactual generated from a normal distribution (Suresh et al. 2017).
5. ***Augmented Feature Occlusion (AFO):*** Augments the Feature Occlusion method by sampling counterfactuals from the bootstrapped distribution over each feature, avoiding generating out-of-distribution samples (Suresh et al. 2017).
6. ***Deep Lift (DL):*** Deep Learning Important FeaTures (Shrikumar, Greenside, and Kundaje 2017) method decomposes the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input.
7. ***Integrated Gradients (IG):*** Assigns This method (Sundararajan, Taly, and Yan 2017) assigns an importance score to each input feature by approximating the integral of gradients of the model's output to the inputs along the path (straight line) from given baselines/references to inputs.
8. ***Gradient Shap (GS):*** Uses the gradient of the model predictions to input features to generate importance scores (Erion et al. 2019).

4 Datasets

In this section, we describe three datasets used in our experiments and their respective initial data processing steps. We compile a new **COVID-19 dataset** dataset, on which we perform both our proposed window-based time series interpretation and sensitivity analysis. Furthermore, to answer our research question: **Is our proposed window-based time series interpretation framework applicable to other models and datasets?** Our proposed workflow is model-agnostic and generic, hence can be applied to any other time series models and datasets to interpret input-output relevance. We answer this by choosing two well known time series datasets: **Electricity** and **Traffic**.

- ***COVID-19 dataset***: Our data is collected from multiple public sources from March 1, 2020, to Dec 29, 2022 (around 3 years) for each of 3,142 US counties. Additionally, we collected weekly COVID-19 cases by age groups from (Centers for Disease Control and Prevention 2023a) to evaluate the age group sensitivity interpretation. These age groups are categorized by the US county population statistics (US Census Bureau 2020) and cases by age groups from CDC (Centers for Disease Control and Prevention 2023a).

  We removed outliers from the data using the following thresholds:

$$\begin{aligned} \text{lower} &= Q1 - (7.5 * IQR) \\ \text{upper} &= Q3 + (7.5 * IQR) \end{aligned} \tag{5}$$

where $Q1$ and $Q3$ represent the first and third percentiles on a weekly moving average basis, and IQR is the interquartile range. We linearly interpolated the missing values and standard normalized the features before training the model. The model uses the previous 2 weeks of data to predict cases for the next 2 weeks.

Feature	Description	Type
Age groups	% of people in each of the 8 age groups	Static
Vaccination	% of fully vaccinated population	Dynamic
Cases	Past infection cases	Dynamic
Month	Timestamps	Known Future
Day in month	Timestamps	Known Future
Day in week	Timestamps	Known Future
Cases	Future infection cases	Target

Table 1: Description of the dataset. Data is collected for each of the 3,142 US counties.

- ***Electricity dataset***: The UCI Electricity Load Diagrams dataset contains the electricity consumption of 321 customers from 2012 to 2014. They are aggregated to an hourly level and normalized. We use the past 96 hours of inputs to forecast for the next 24 hours. We also added four time-encoded features: month, day of the month, day of the week, and hour. Following (Zhang and Yan 2022) we use the record of customer 'MT_321' as the time series of interest.

Feature	Description	Type
Consumption	Past electricity consumption	Dynamic
Month	Timestamps	Known Future
Day in month	Timestamps	Known Future
Day in week	Timestamps	Known Future
Hour	Timestamps	Known Future
Consumption	Future electricity consumption	Target

Table 2: Description of Electricity dataset.

- ***Traffic dataset***: The UCI PEM-SF Traffic dataset describes the occupancy rate (with $y_t \in [0, 1]$) of 440 SF Bay Area freeways from 2015 to 2016. We perform the same data processing steps with the Electricity dataset. Following (Zhang and Yan 2022) we use the record of 821st station user as the time series of interest.

5 Experiment Setup

This section describes our experimental setup for preparing the best time series model for the infection forecasting task at the daily US county level. In the next section, we describe how we interpret this model using black-box interpretation methods. All of our experiments were done in a remote HPC server with NVIDIA V100 GPU and 32 GB memory.

Feature	Description	Type
Occupancy rate	Past road occupancy rate	Dynamic
Month	Timestamps	Known Future
Day in month	Timestamps	Known Future
Day in week	Timestamps	Known Future
Hour	Timestamps	Known Future
Occupancy rate	Future road occupancy rate	Target

Table 3: Description of Traffic dataset.

5.1 COVID-19 Data Preprocessing

The data set was split into training, validation, and testing sets. The training set includes March 1, 2020, through November 27, 2021. The immediate next 2 weeks are used as the validation set and the next 2 weeks after that are used as the test set. The best model checkpointed by the validation set is loaded later for testing. We use additional data after this period for deployment benchmark in Section 3.2.

5.2 Models and Parameters

We use the following models for benchmarking: TimesNet (Wu et al. 2023a), PatchTST (Nie et al. 2023), FEDformer (Zhou et al. 2022), Autoformer (Wu et al. 2021), Crossformer (Zhang and Yan 2022). We implement these models [1] using The Time-Series-Library (Wu et al. 2023b).

The model hyper-parameters were chosen based on (Wu et al. 2023a) and tuning. Table 4 reports the common parameters used by the selected models during the experiment. Full documentation can be found in our project repository.

Parameter	Value	Parameter	Value
learning rate	1e-3	loss	MSE
batch size	32	dropout	0.1
encoder layers	2	random seed	7
decoder layers	1	hidden size	64
attention heads	4	label length	7

Table 4: Common model hyperparameters.

5.3 Implementation Library

We used the Captum (Kokhlikyan et al. 2020) and Time Interpret (Enguehard 2023) libraries to implement these interpretation methods. Except for the Morris Sensitivity (Morris 1991), which was implemented using the Sensitivity Analysis Library (Iwanaga, Usher, and Herman 2022). The baselines to mask the input features were randomly generated from a normal distribution. Unlike (Enguehard 2023), which runs the interpretation on CPU, our implementation modifies the prior libraries to work on GPU. For interpretation methods requiring a bootstrapped distribution for baseline generation, we used the training data as the distribution.

[1] https://github.com/UVA-MLSys/COVID-19-age-groups

5.4 Prediction Results

We use the following popular evaluation metrics for our regression task: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Root Mean Squared Logarithmic Error (RMSLE), and Coefficient of determination (R^2-score). For all metrics, except R^2-score, the lower the error loss the better. For R^2-score, 1.0 is the best possible score and it can be negative if the model is arbitrarily worse.

Model	MAE	RMSE	RMSLE	R^2-score
Autoformer	35.69	189.4	1.918	0.451
FEDformer	**30.19**	**182.2**	**1.467**	**0.481**
PatchTST	31.17	183.6	1.530	0.469
TimesNet	34.35	191.9	1.604	0.415
Crossformer	39.58	193.6	2.141	0.394

Table 5: Test performance of the deep learning models. The best results are in bold.

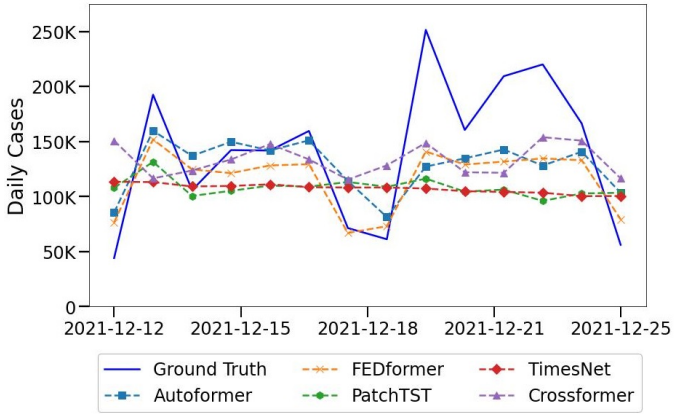

Figure 1: Test predictions comparison with ground truth aggregated over all counties.

Table 5 shows the test results of the models. The FEDformer model performs best with the lowest MAE, RMSE, RMSLE loss, and highest R^2-score. We have used this FEDformer model in our later section for the interpretation tasks. Figure 1 plots the ground truth and model predictions aggregated over the 3,142 US counties. The aggregated plot is shown for simplicity. However, the prediction and evaluation are done at each US county level.

5.5 Two Additional Experimental Setup

The Electricity and Traffic datasets are divided into train, validation test sets using a 8:1:1 split. We arbitrarily chose Crossformer for these two datasets. The MAE and MSE errors on the test are 0.2534 and 0.1292 for the Electricity dataset. For the Traffic dataset, the MAE and MSE errors are 0.2938 and 0.2258 respectively.

5.6 Interpretation Evaluation

Following the same steps as described in Section 3.1, we apply all of our eight interpretation methods on our **COVID** dataset and display our results in Table 6. Furthermore, we showcase the performance of an arbitrarily chosen subset of interpretation methods in Table 7 and 8 for the Electricity and Traffic datasets.

Method	Comprehensiveness (↑)		Sufficiency(↓)	
	MAE	MSE	MAE	MSE
Feature Ablation	4.91	8.64	9.53	10.5
Feature Permutation	4.00	7.08	8.00	8.28
Morris Sensitivity	**6.23**	9.39	**5.85**	**5.46**
Feature Occlusion	4.89	8.44	9.49	10.4
Augmented F.O.	4.18	7.66	7.96	8.09
Deep Lift	5.72	**9.54**	8.90	9.43
Integrated Gradients	5.52	9.09	9.25	10.2
Gradient Shap	4.78	8.17	8.04	8.27

Table 6: AOPCR results of the interpretation using the FEDformer model on the **COVID** test set.

Method	Comprehensiveness (↑)		Sufficiency(↓)	
	MAE	MSE	MAE	MSE
Feature Ablation	**13.4**	**12.2**	17.0	18.1
Feature Permutation	7.57	5.28	**15.2**	**14.8**
Feature Occlusion	13.3	12.2	17.1	18.4
Augmented F.O.	8.27	6.12	15.3	15.1

Table 7: AOPCR results of the interpretation using the Crossformer model on the **Electricity** test set.

Method	Comprehensiveness (↑)		Sufficiency(↓)	
	MAE	MSE	MAE	MSE
Feature Ablation	10.8	7.70	**16.3**	**16.6**
Feature Permutation	8.20	4.72	19.6	22.1
Feature Occlusion	**10.9**	**7.76**	16.4	16.8
Augmented F.O.	8.39	4.89	19.4	21.8

Table 8: AOPCR results of the interpretation using the Crossformer model on the **Traffic** test set.

6 Evaluating Age Group Sensitivity with Ground Truth

This section presents an innovative way to evaluate the importance of age groups from CDC (Centers for Disease Control and Prevention 2023b). The previous section uses performance drop-based methods to evaluate interpretation because of the lack of ground truth for interpretation. And that evaluation is done at each county and daily level, the same as the prediction task.

However, the COVID-19 cases by age groups from CDC (Centers for Disease Control and Prevention 2023b) comes at a weekly rate for the whole United States. Hence can only be evaluated at the weekly and aggregated to the country level. Figure 2 shows the ground truth values by age groups over our whole dataset. Table 9 shows the summary and rank of different age groups for the test period.

The weekly rank of the age groups by infection rate doesn't change much with time. However, the infection rate

Age Group	Total Cases		
	Actual	Normalized (%)	Rank
< 5	99654	3.569	7
5-17	404420	14.48	4
18-29	686648	24.59	1
30-39	539684	19.32	2
40-49	393727	14.10	5
50-64	443701	15.89	3
65-74	141490	5.067	6
75+	83086	2.975	8

Table 9: COVID-19 cases in all US counties by age groups during the test period, 12-25 Dec 2021.

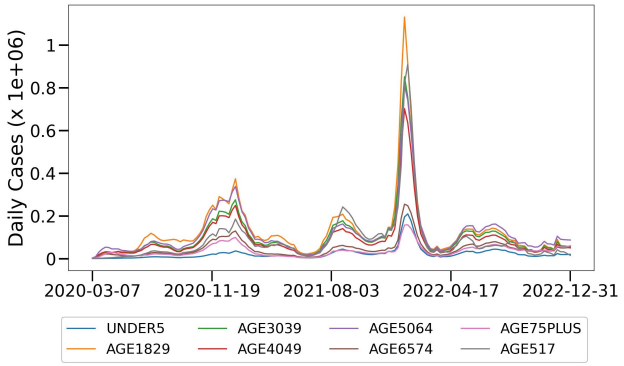

Figure 2: Weekly COVID-19 cases (Centers for Disease Control and Prevention 2023a) for each of the eight age subgroups over the study period .

itself can vary a lot and be more challenging. Hence we focus on predicting the normalized (l1-norm) infection rate for each age group. The normalization is done to understand how each age group contributes to the overall infection spread each week.

6.1 Evaluation on Extended Dataset

Our test set initially comprises only two weeks of data, but we extend the evaluation until December 31, 2022, encompassing over a year's worth of test data. This extension aims to demonstrate the alignment of our predicted age sensitivity with the actual cases reported by the CDC for the United States. Figure 3 compares the extended dataset. Both predicted attribution from the importance matrix ϕ and actual sensitivity (cases by age groups) are normalized to sum to 1.00. The results show that the trends of different age groups keep changing with time.

6.2 Interpreting different feature attribution

Table 10 shows the aggregated importance of the input features over the test set using the Morris Sensitivity method. We average the attribution matrix ϕ and normalize the scores to percentages. The past COVID-19 cases are the most important. Among other features, the age groups 18-29 and 65-75 are more important.

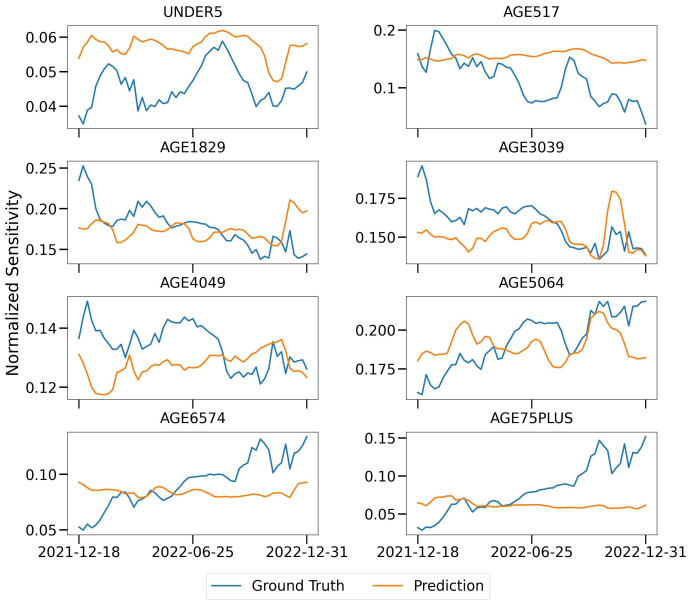

Figure 3: Predicted age sensitivity on the extended dataset with weekly COVID-19 cases by age groups as ground truth.

An interpretation example is shown in Figure 4 for the Los Angeles, California county from our test period. The position index from -14 to -1 is for the input, while position 0 to 13 is for the prediction horizon. The result shows that the same feature from different lookback positions has different impact on the predictions. The past cases are most important and working-age populations (AGE3039 and AGE4049) also get higher importance. Recent cases get higher attribution values, showing the model's prediction is more relevant to recent infection rates.

We calculated how much this normalized infection rate differs from the predicted scores using different interpretation methods. The importance matrix $\phi \in \mathbf{R}^{\tau_{max} \times J \times L}$ is aggregated over the lookback window L to find the overall impact of each age feature j for the prediction horizon τ. The difference between the actual and the predicted rate through interpretation is evaluated using MAE, RMSE, and NDCG (Normalized Discounted Cumulative Gain). The NDCG ranking metric returns a high value if true labels are

Feature	Importance	Feature	Importance
UNDER5	4.737	AGE75PLUS	4.695
AGE517	4.785	Vaccination	4.385
AGE1829	5.264	Cases	**41.98**
AGE3039	4.946	Day	4.961
AGE4049	5.191	Month	4.611
AGE5064	4.594	Weekday	4.594
AGE6575	5.258		

Table 10: Feature importance (%) evaluated on the test set by aggregating attribution scores for the input features.

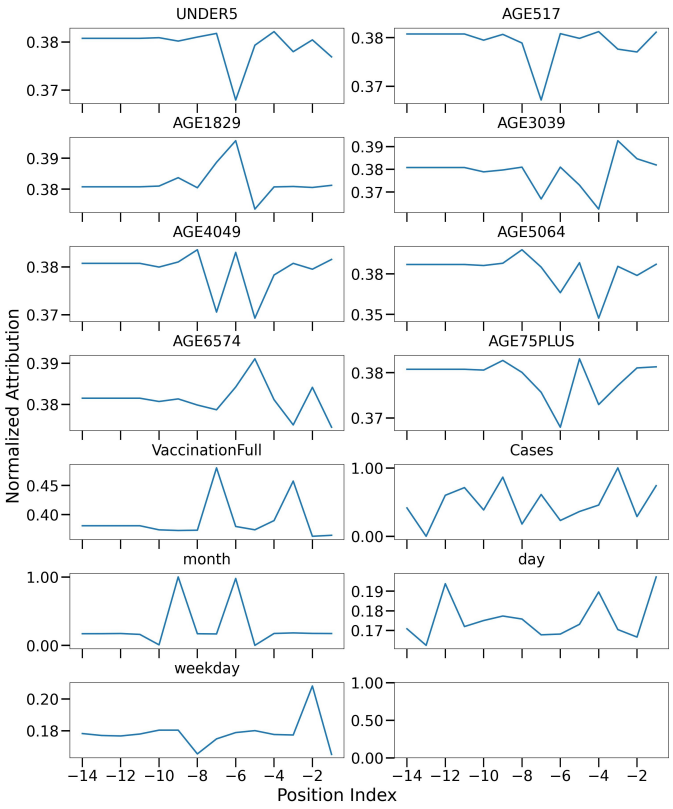

Figure 4: Interpreting different feature attribution for the Los Angeles, California county from our test set using the FEDformer model and Feature Ablation method.

ranked high by the predicted scores.

Table 11 shows the final results. There is no single best method from the results but most methods achieve high accuracy in predicting the age group sensitivities.

7 Discussion

We discuss the time complexity of our approach and the code reproducibility of our work.

7.1 Time Complexity

Execution time is important for real-time applications. We report the execution time of our interpretation methods and experiments on the test set in Table 12. We observe the gradients-based methods perform faster. The Morris Sensitivity method is slower due to multiple sampling of the input features.

7.2 Code Reproducibility

Our code and datasets are publicly available on GitHub at https://github.com/UVA-MLSys/COVID-19-age-groups. We have also published a singularity container documenting the software versions. This also helps to readily deploy it on any HPC cluster. Our random methods are seeded to ensure reproducibility.

Method	Metrics		
	MAE ↓	RMSE ↓	NDCG ↑
Feature Ablation	**0.0336**	0.0426	0.9849
Feature Permutation	0.0339	0.0438	**0.9900**
Morris Sensitivity	0.0350	**0.0393**	0.9587
Feature Occlusion	0.0338	0.0429	0.9851
Augmented F.O.	0.0346	0.0446	0.9587
Deep Lift	0.0383	0.0455	0.9641
Integrated Gradients	0.0386	0.0460	0.9641
Gradient Shap	0.0336	0.0431	0.9713

Table 11: Evaluation of predicted normalized attribution scores with normalized weekly **COVID-19 cases** by age group in the test period, 12-25 Dec 2021. The best results are in bold.

Method	Time	Method	Time
Morris Sensitivity	294.1	Feature Ablation	162.9
Feature Permutation	162.8	Augmented F.O.	165.3
Feature Occlusion	165.3	Deep Lift	55.40
Integrated Gradients	120.4	Gradient Shap	59.58

Table 12: Execution time (seconds) of the interpretation methods on the test set.

8 Conclusion and Future Work

In this work, we showed how to interpret state-of-the-art time series transformer models using very dynamic and complex COVID-19 infection data and two benchmark time series datasets using Electricity and Traffic. We provide a thorough analysis of recent times series interpretation methods performance on the state-of-the-art Transformer models. Our results show that we can not only interpret changes in feature importance over past time steps but also predict the sensitivity of these features in future horizons. Our proposed framework helps us understand the impacts of past observations, but also predict their impacts in the future. Future works include capturing higher-order relations between the input features, understanding spatiotemporal interpretations better, and benchmarking more time series domains with our framework.

9 Acknowledgment

This work is partially supported by NSF grant CCF-1918626 Expeditions: Collaborative Research: Global Pervasive Computational Epidemiology, and NSF Grant 1835631 for CINES: A Scalable Cyberinfrastructure for Sustained Innovation in Network Engineering and Science.

References

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

Arik, S. O.; Yoder, N. C.; and Pfister, T. 2022. Self-Adaptive

Forecasting for Improved Deep Learning on Non-Stationary Time-Series. *arXiv preprint arXiv:2202.02403*.

Centers for Disease Control and Prevention. 2023a. COVID-19 Weekly Cases and Deaths by Age, Race/Ethnicity, and Sex.

Centers for Disease Control and Prevention. 2023b. COVID-19 Weekly Cases and Deaths by Age, Race/Ethnicity, and Sex.

Clement, J. C.; Ponnusamy, V.; Sriharipriya, K.; and Nandakumar, R. 2021. A survey on mathematical, machine learning and deep learning models for COVID-19 transmission and diagnosis. *IEEE reviews in biomedical engineering*, 15: 325–340.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2019. ERASER: A benchmark to evaluate rationalized NLP models. *arXiv preprint arXiv:1911.03429*.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Enguehard, J. 2023. Time Interpret: a Unified Model Interpretability Library for Time Series. *arXiv preprint arXiv:2306.02968*.

Erion, G. G.; Janizek, J. D.; Sturmfels, P.; Lundberg, S. M.; and Lee, S.-I. 2019. Learning Explainable Models Using Attribution Priors. *ArXiv*, abs/1906.10670.

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.

Iwanaga, T.; Usher, W.; and Herman, J. 2022. Toward SALib 2.0: Advancing the accessibility and interpretability of global sensitivity analyses. *Socio-Environmental Systems Modelling*, 4: 18155.

Kim, D.; Min, H.; Nam, Y.; Song, H.; Yoon, S.; Kim, M.; and Lee, J.-G. 2022. Covid-eenet: Predicting fine-grained impact of COVID-19 on local economies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11971–11981.

Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; and Reblitz-Richardson, O. 2020. Captum: A unified and generic model interpretability library for PyTorch. arXiv:2009.07896.

Molnar, C. 2020. *Interpretable machine learning*. Lulu. com.

Morris, M. 1991. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2): 161–174.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A time series is worth 64 words: Long-term forecasting with transformers. *International Conference on Learning Representations*.

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

Ramchandani, A.; Fan, C.; and Mostafavi, A. 2020. Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions. *Ieee Access*, 8: 159915–159930.

Rodriguez, A.; Tabassum, A.; Cui, J.; Xie, J.; Ho, J.; Agarwal, P.; Adhikari, B.; and Prakash, B. A. 2021. Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15393–15400.

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*.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, 3145–3153. PMLR.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *International Conference on Machine Learning*.

Suresh, H.; Hunt, N.; Johnson, A. E. W.; Celi, L. A.; Szolovits, P.; and Ghassemi, M. 2017. Clinical Intervention Prediction and Understanding using Deep Networks. *ArXiv*, abs/1705.08498.

Turbé, H.; Bjelogrlic, M.; Lovis, C.; and Mengaldo, G. 2023. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*, 5(3): 250–260.

US Census Bureau. 2020. County Population by Characteristics: 2010-2020.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023a. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023b. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *Advances in Neural Information Processing Systems*.

Zeiler, M. D.; and Fergus, R. 2013. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision*.

Zhang, Y.; and Yan, J. 2022. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*.

Zhao, Z.; Shi, Y.; Wu, S.; Yang, F.; Song, W.; and Liu, N. 2023. Interpretation of Time-Series Deep Models: A Survey. *arXiv preprint arXiv:2305.14582*.

Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proc. 39th International Conference on Machine Learning (ICML 2022)*.