NTNU

Norwegian University of Science and Technology

IARIA

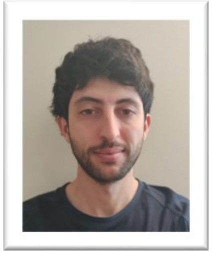
# An XAI Approach on the Capacity of Transformers to Learn Time Dependencies in Time Series Forecasting

Authors: Alberto Miño Calero [1], Adil Rasheed [1], Anastasios M. Lekkas [1]

Presenter: Alberto Miño Calero

[1] Norwegian University of Science and Technology, Norway

Contact email: alberto.m.calero@ntnu.no



# Alberto Miño Calero

- Holds a bachelor's degree in Computer Science specializing in Software Engineering from the University of Cordoba and a Master's degree in Computer Science and Technology from the University Carlos III of Madrid.
- Currently a Ph.D. candidate in the Department of Engineering Cybernetics of the NTNU studying explainable AI in the context of Deep Learning.
- His research focuses on improving our understanding of the behavior of neural networks to gain insight on the patterns they learn to solve tasks and compare the use of simpler model against Deep Learning looking at features such as reliability, stability, and robustness.

# Goal and contributions

Goal:

Assess the suitability of transformers in time series forecasting tasks by investigating what time dependencies they can learn.

Contributions:

We propose **methodology to analyze time dependencies** learned by transformers **based on Shapley additive explanations**.

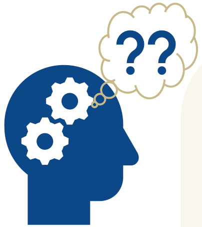
We investigate a variety of **aggregation strategies** with the input time series **to visualize these time dependencies**.

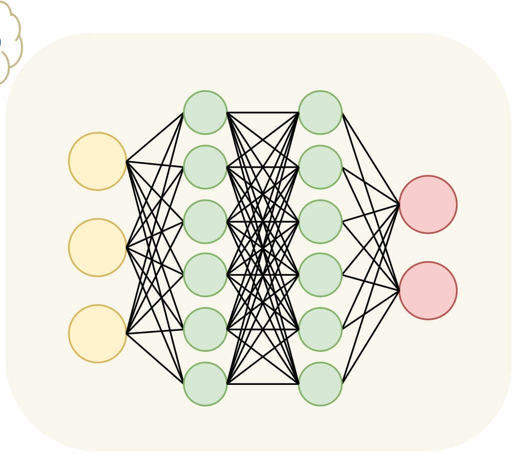
We find the **transformer is unable to learn long-term time dependencies and just looks at** the very end of the input sequence.

# Introduction: time series forecasting

- Time series forecasting (LTSF) problems play a major role in many research [1] and applied domains [2 – 5]:
  - Climate, healthcare, biology, economics, physics…

- Solving with deep learning means giving up interpreting the solutions.
  - Are they working as intended?





- Many deep learning approaches:
  - Multilayer perceptron [6], convolutional NNs [7].
  - Recurrent [8], long short-term memory networks [9].
  - Transformers [10] have gained a lot of traction [11] – [15]:
    - High performance with sequential data.
    - Handling of contextual information.

# Introduction: time series forecasting

## Challenge of Deep Learning

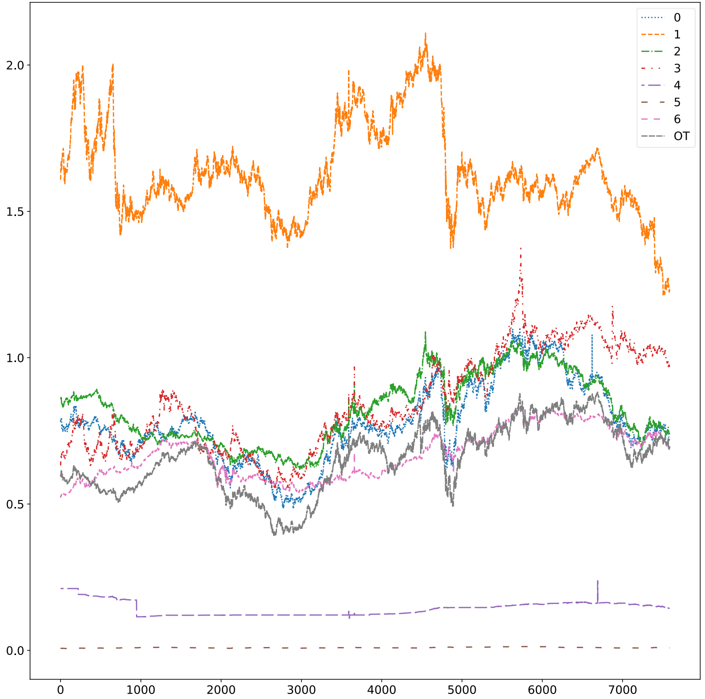
- Neural networks are black boxes:
  - Not interpretable.
  - We do not understand their predictions.
- XAI methods can provide this knowledge:
  - Exploiting external [16] or internal elements [17], [18].
  - Model agnostic [16] or model specific [17], [18].
- Popular XAI methods for transformers:
  - Attribution score-based [16] – [18].
  - Attention-based [19] – [21].

## Concern with transformers

- Pitfall in performance:
  - Simple linear autoregressive models (LTSF-Linear) can outperform them [22]:
    - Compared with mean square and mean absolute errors.
- Designed for natural language processing [10]:
  - Contextual information equals time dependencies?
  - Can attention deal with long-term time dependencies?
- XAI can help us understand where they are failing.

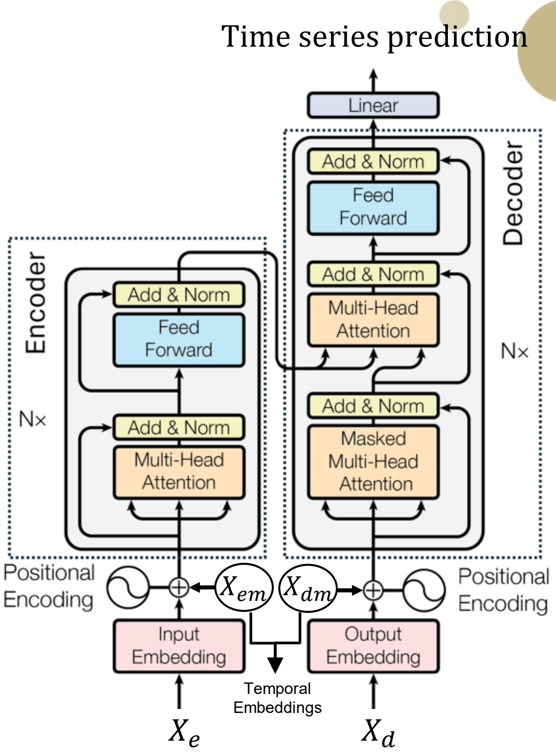
# Methodology: data

- Currency exchange rate dataset:
  - Rates of 8 countries.
  - Collected between 1990 and 2016.
  - Publicly available [22], [23].
  - Used as time series forecasting benchmark of transformer-based models [12], [22].
  - Time resolution of 1 day:
    - Total of 7588 samples.
- Problem setting:
  - Multivariate forecasting.



# Methodology: model

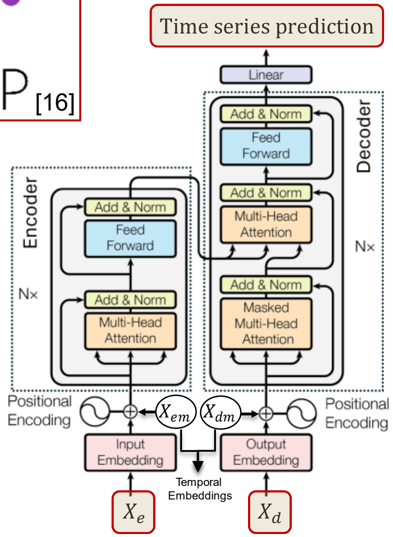
- “Vanilla” transformer, commonly used in LTSF comparisons of transformers performance [12], [22]:
  - Enhanced start token:
    - $L_{dec}$ = 48 (time steps).
  - Includes temporal embeddings for encoder and decoder: $X_{em}$, $X_{dm}$.
  - $X_{em}$, $X_{dm}$, and decoder input $(X_d)$ treated as regular model inputs, together with the input sequence $(X_e)$:
    - Not transparent to the user.
  - Direct multi-step prediction:
    - $Z_p = 96$.
  - Input time series sequence length:
    - $Z_i = 96$.



# Methodology: SHAP

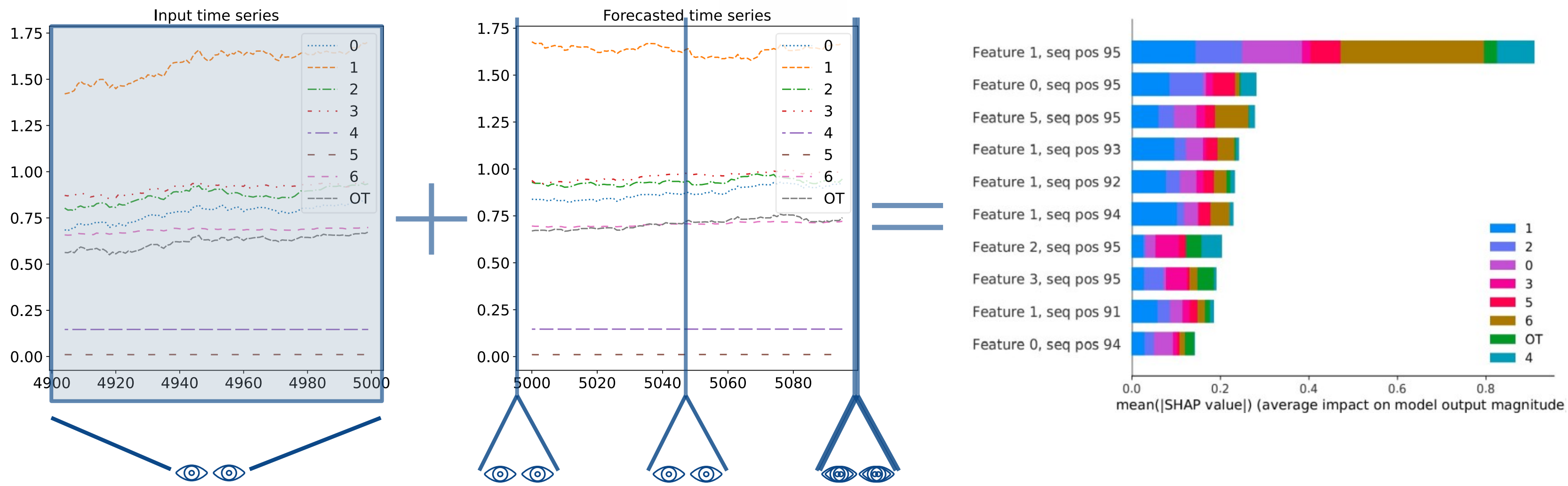
- Shapley additive explanations [16]:
- SHAP advantages:
  - Strong theoretical base:
    - Coalition game theory.
  - Model agnostic.
  - Local explanations.
    - Computes attribution scores for the inputs based on how an output reacts to perturbing the inputs.
    - **We use it to extract global explanations by aggregating local ones**.
  - Several implementations:
    - Choose the best depending on model and dataset: Deep implementation.
  - Package SHAP available:
    - Compatible with Pytorch (and Tensorflow).
- Main disadvantage found in this work:
  - If the model is not supported, meeting the specifications can be a challenge.

SHAP [16]



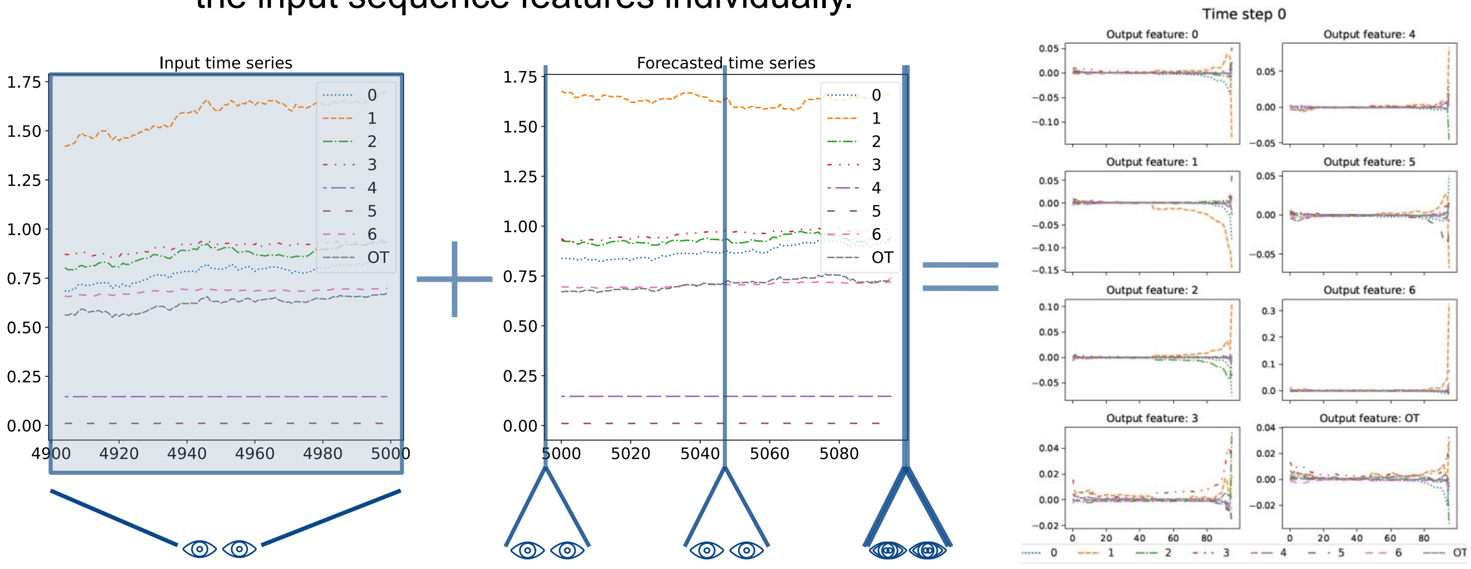
# Methodology: the analysis

- Three levels according how locally or globally the model is described:
  - Local:
    - The 10 most important features looking at specific predicted time steps considering each input feature from each time step independently.



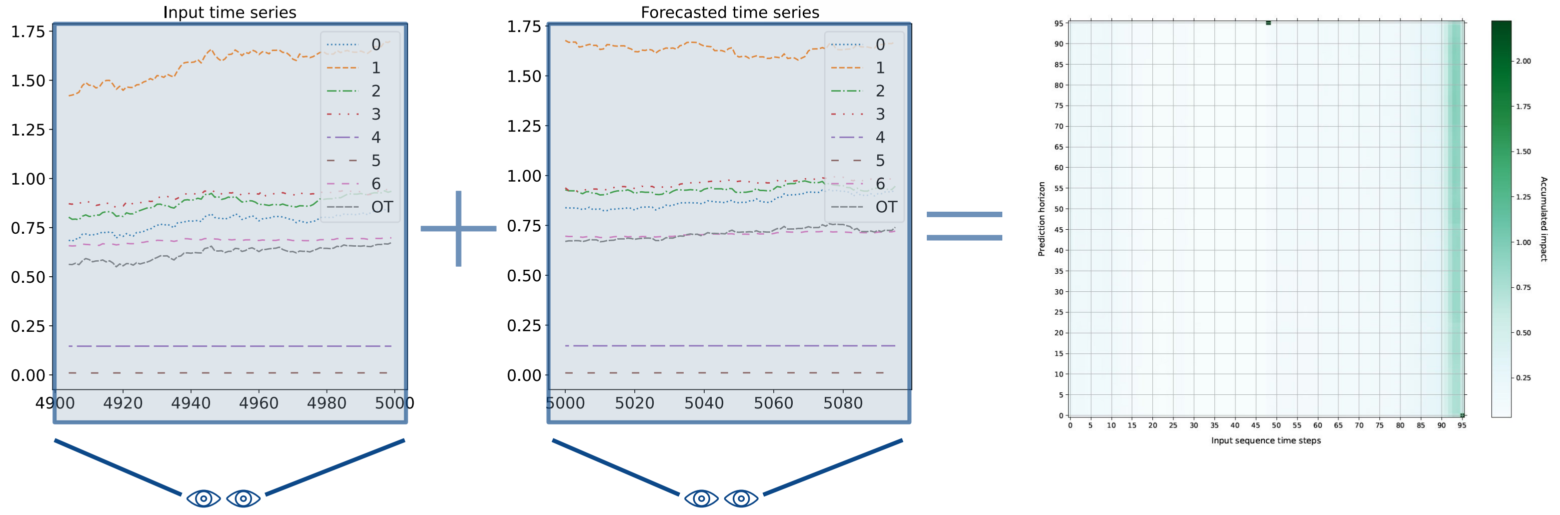
# Methodology: the analysis

- Three levels according how locally or globally the model is described:
  - Intermediate:
    - Evolution of attribution scores computed for several specific predicted time steps for all the input sequence features individually.



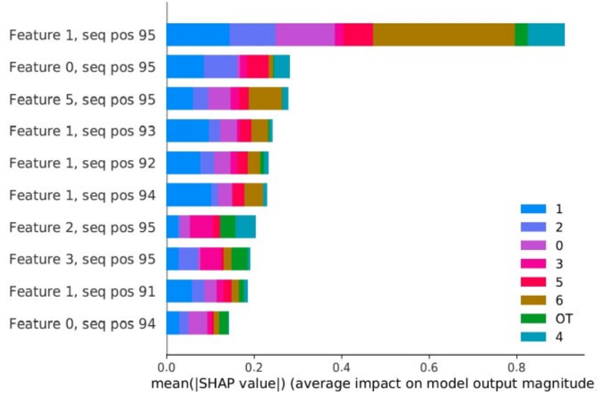
# Methodology: the analysis

- Three levels according how locally or globally the model is described:
  - Global:
    - Evolution of attribution scores for all predicted time steps and all input sequence features, accumulating both input scores and outputs by time step.

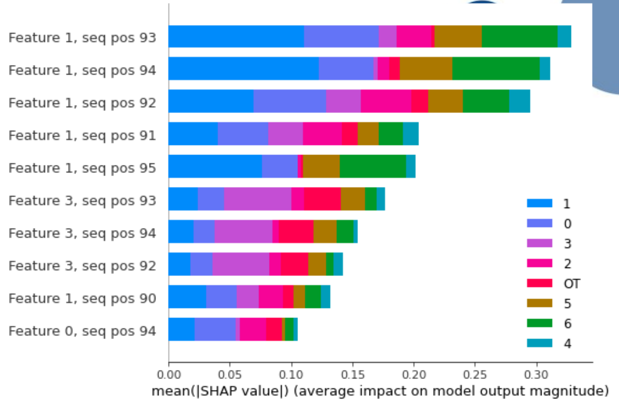


# Results

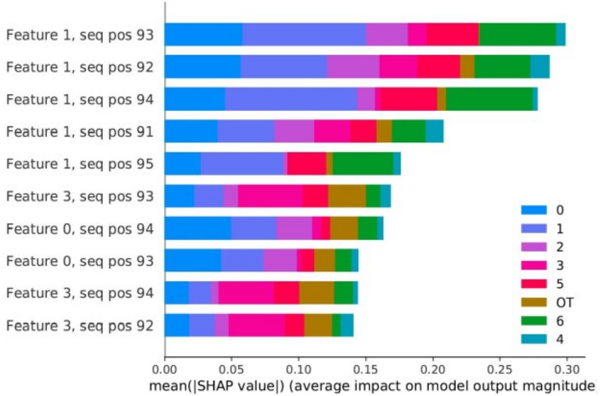
- Last input time steps are always among the 10 most influential features.
- Behavior consistent in all the forecasted for all time steps, expect:
    - Anomaly at $Z_p = 95$.
    - No clear reason.
    - Nothing of particular interest in the dataset.
    - Completely different from $Z_p = 94$.
    - Most likely an outlier in prediction.
- No evident presence of learned long-term time dependencies.



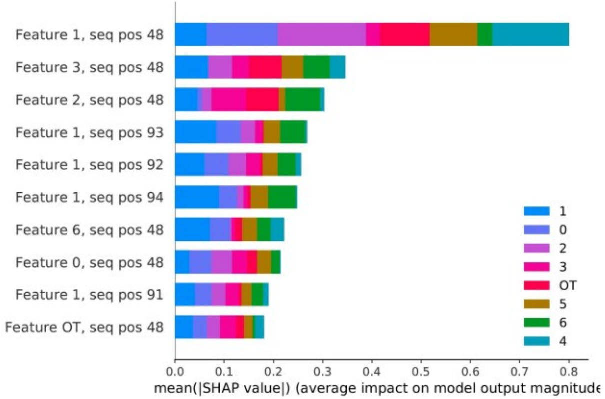
$Z_p = 0$



$Z_p = 48$



$Z_p = 94$

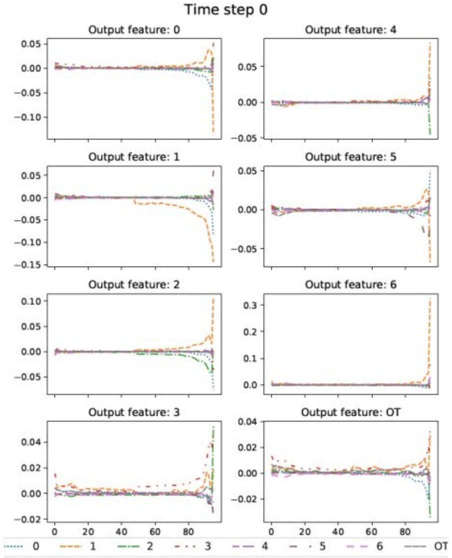


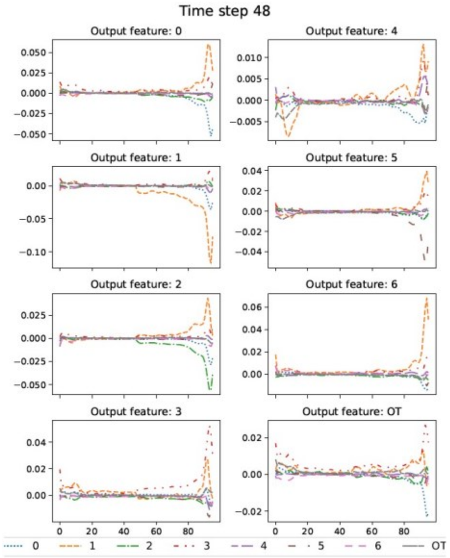
$Z_p = 95$

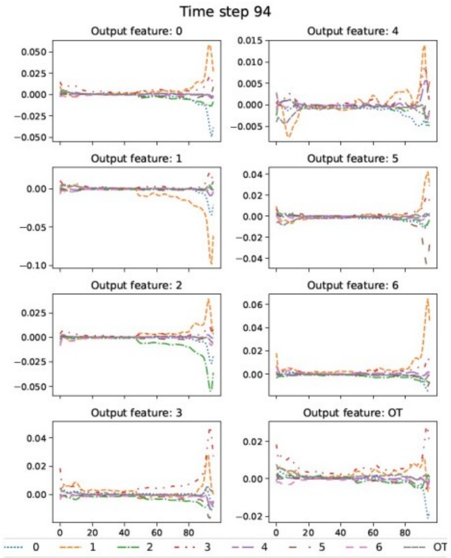
# Results

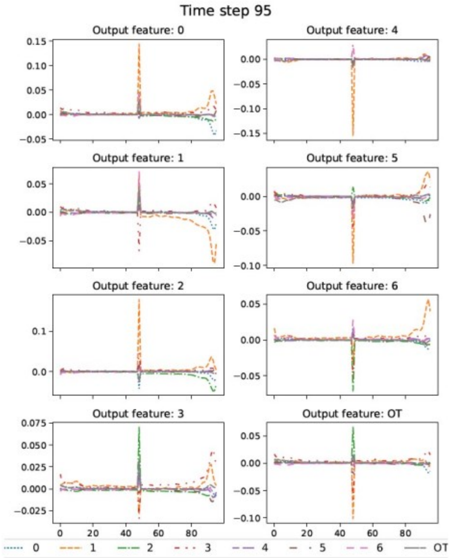
- Feature importance grows as we move towards the last time steps of the input.
- After few first forecasted time steps, the impact is more distributed.
  - Some input features from beginning and middle input time steps have observable impact.

- Most of the impact is still placed in the last time steps of the input sequence.
- The anomaly can be better seen:
  - Input time step $Z_i = 48$ has an anomalous impact on $Z_p = 95$.



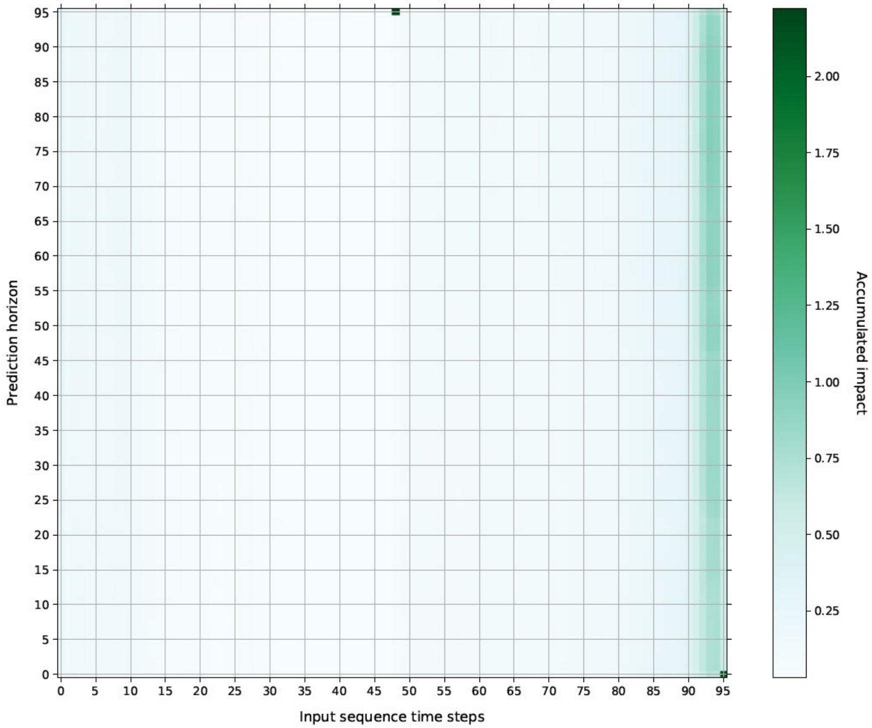






# Results

- Most of the influence is place on very few time steps:
  - The output is mostly affected by the last five time steps of the input sequence.
  - This suggest the model is unable to learn any long-term time dependencies.
- There are two anomalies:
  - First at $Z_p = 0$:
    - Just in terms of accumulated impact value, behavior still consistent.
  - Second at $Z_p = 95$:
    - The same spotted in previous figures.
    - This only anomaly in behavior from 9216 datapoints suggest an outlier in prediction from the side of the transformer.



# Conclusions and future work

## Conclusions:

- We propose a methodology to analyze transformers with SHAP in the LTSF domain.
- We find the transformer does not learn long-term time dependencies:
  - Predictions are mostly influenced by the last elements of the input sequence.
  - The transformer disregards most of the input time series in this case.

## Future work:

- More extensive analysis on different datasets and with state-of-the-art transformers designed for LTSF.
- More in-depth analysis when new transformers are proposed in this domain could be useful to detect these issues, instead of just looking at performance.

# References

[1] G. Mahalakshmi, S. Sridevi, and S. Rajaram, “A survey on forecasting of time series data,” in 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE’16), Jan. 2016, pp. 1–8. DOI: 10.1109/ICCTIDE.2016.7725358.

[2] M. Mudelsee, Climate time series analysis. Classical statistical and bootstrap methods. (Atmospheric and Oceanographic Sciences Library), Second Edition. Springer, 2014, vol. 51, ISBN: 978-3-319-04449-1. DOI: 10.1007/978-3-319-04450-7.

[3] E. J. Topol, “High-performance medicine: The convergence of human and artificial intelligence,” Nature Medicine, vol. 25, no. 1, pp. 44–56, Jan. 2019, Publisher: Nature Publishing Group, ISSN: 1546-170X. DOI: 10.1038/s41591-018-0300-7.

[4] B. Lim and S. Zohren, “Time-series forecasting with deep learning: A survey,” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 379, no. 2194, p. 20 200 209, Feb. 2021, Publisher: Royal Society. DOI: 10.1098/rsta.2020.0209.

[5] J. Feyrer, “Trade and Income—Exploiting Time Series in Geography,” American Economic Journal: Applied Economics, vol. 11, no. 4, pp. 1–35, Oct. 2019, ISSN: 1945-7782. DOI: 10.1257/app.20170616.

[6] T. Kolarik and G. Rudorfer, “Time series forecasting using neural networks,” ACM SIGAPL APL Quote Quad, vol. 25, no. 1, pp. 86–94, 1994, ISSN: 0163-6006. DOI: 10 . 1145 / 190468.190290.

[7] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, “Convolutional neural networks for time series classification,” Journal of Systems Engineering and Electronics, vol. 28, no. 1, pp. 162–169, Feb. 2017, Conference Name: Journal of Systems Engineering and Electronics, ISSN: 1004-4132. DOI: 10.21629/JSEE.2017. 01.18.

[8] H. Hewamalage, C. Bergmeir, and K. Bandara, “Recurrent Neural Networks for Time Series Forecasting: Current status and future directions,” International Journal of Forecasting, vol. 37, no. 1, pp. 388–427, Jan. 2021, ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2020.06.008.

[9] U. M. Sirisha, M. C. Belavagi, and G. Attigeri, “Profit Prediction Using ARIMA, SARIMA and LSTM Models in Time Series Forecasting: A Comparison,” IEEE Access, vol. 10, pp. 124 715–124 727, 2022, Conference Name: IEEE Access, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3224938.

[10] A. Vaswani et al., “Attention is All you Need,” in Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., Jun. 2017, p. 11.

[11] H. Zhou et al., “Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting,” in Proceedings of the AAAI Conference on Artificial Intelligence, ISSN: 2374-3468, 2159-5399 Issue: 12 Journal Abbreviation: AAAI, vol. 35, May 2021, pp. 11 106–11 115. DOI: 10 . 1609 / aaai . v35i12.17325.

[12] H. Wu, J. Xu, J. Wang, and M. Long, Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting, arXiv:2106.13008 [cs], Jan. 2022. DOI: 10.48550/arXiv.2106.13008.

# References

[13] G. Woo, C. Liu, D. Sahoo, A. Kumar, and S. Hoi, ETSformer: Exponential Smoothing Transformers for Time-series Forecasting, arXiv:2202.01381 [cs], Jun. 2022. DOI: 10.48550/arXiv. 2202.01381.

[14] Y. Zhang and J. Yan, “Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting,” in ICLR Proceedings 2023, Sep. 2022, p. 21.

[15] Q. Wen et al., Transformers in Time Series: A Survey, arXiv:2202.07125 [cs, eess, stat], May 2023.

[16] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017.

[17] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in Proceedings of the 34th International Conference on Machine Learning - Volume 70, ser. ICML’17, Sydney, NSW, Australia: JMLR.org, 2017, pp. 3319–3328.

[18] S. Bach et al., “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation,” en, PLOS ONE, vol. 10, no. 7, e0130140, Jul. 2015, Publisher: Public Library of Science, ISSN: 1932-6203. DOI: 10.1371/ journal.pone.0130140.

[19] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, “Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5797–5808. DOI: 10.18653/v1/P19-1580.

[20] S. Abnar and W. Zuidema, “Quantifying Attention Flow in Transformers,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 4190–4197. DOI: 10.18653/v1/2020.acl-main.385.

[21] H. Chefer, S. Gur, and L. Wolf, Transformer Interpretability Beyond Attention Visualization, en, arXiv:2012.09838 [cs], Apr. 2021.

[22] A. Zeng, M. Chen, L. Zhang, and Q. Xu, “Are Transformers Effective for Time Series Forecasting?” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 9, pp. 11 121–11 128, Jun. 2023, Number: 9, ISSN: 2374-3468. DOI: 10.1609/aaai.v37i9.26317.

[23] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, “Modeling Longand Short-Term Temporal Patterns with Deep Neural Networks,” in The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, ser. SIGIR ’18, New York, NY, USA: Association for Computing Machinery, Jun. 2018, pp. 95–104, ISBN: 978-1-4503-5657-2. DOI: 10.1145/3209978.3210006.