Temporal Attention Signatures for Interpretable Time-Series Prediction

Alexander Katrompas and Vangelis Metsis

Texas State University, San Marcos TX 78666, USA
amk181@txstate.edu, vmetsis@txstate.edu

Abstract. Deep neural networks have become a staple in time-series prediction due to their remarkable accuracy. However, their internal workings often remain elusive. Significant advancements have been made in the interpretability of these networks, with attention mechanisms and feature maps being notably effective for image classification by highlighting the crucial data points. While human observers can readily confirm the significance of features in image classification, the interpretability of time-series data and its modeling remains challenging. To address this, we put forth an innovative approach that unifies temporal attention and visualization as a blend of recurrent neural networks, self-attention, and general attention. This synergy results in the generation of *temporal attention signatures*, akin to image attention heat maps. Temporal attention not only enhances prediction accuracy beyond that of recurrent networks alone but also demonstrates that varying label classes yield distinct attention signatures. This observation indicates that neural networks focus on different sections of time-series sequences contingent on the prediction target. We conclude with a discussion on the practical implications of this novel approach, including its applicability to model interpretation, sequence length selection, and model validation. This leads to more accurate, robust, and interpretable models, instilling greater confidence in their results.

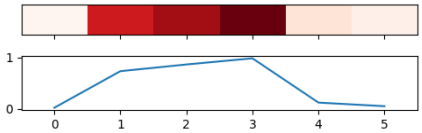
Keywords: Neural Networks · Deep Learning · Attention Mechanisms · Time-Series · Model Interpretability.

1 Introduction

Recurrent neural network models (RNNs) have a long, successful history in time-series modeling [13]. Recent advancements in the combination of attention mechanisms with Long Short-Term Memory (LSTM) networks, a type of RNN, have demonstrated that LSTMs can achieve improved performance, surpassing both RNNs and attention-only models (i.e., transformers) [9,20]. As a result, RNNs, particularly LSTMs, remain competitive for modeling complex time-series data, particularly in hybrid models [13,20].

Attention mechanisms have been successfully used in text processing [3,11,18] and in image classification. Attention has also led to advances in model interpretability, wherein image attention heat maps validate the network's attention [7,10]. Inspired by both the accuracy of RNN-attention models and image

(a) Visual attention heatmap [1].



(b) Temporal attention signature sequence length 6.

Fig. 1: Examples of self-attention visualization demonstrating which parts of a sequence are important to the classification decision.

attention interpretability, we introduce a novel RNN-attention time-series model, which we refer to as *temporal attention*, used to achieve state-of-the-art accuracy and interpretability via *temporal attention signatures*.

For comparison to image classification, Figure 1a shows an image attention heat map, which reveals what the classification network identifies as significant [1]. The visualization is easily validated through observation, providing credibility to the model [1]. Time-series data and models are typically more challenging to interpret. Therefore, despite their accuracy, time-series models may be viewed with some skepticism [16]. In the work, we develop temporal attention signatures, providing a method of interpretation and validation for time-series models which is conceptually comparable but computationally innovative. For illustration, Figure 1b shows an attention signature displaying a six-step time-series sequence, wherein it can easily be seen that the middle of the sequence significantly influences the prediction of the specific label.

Temporal attention signatures demonstrate that the network consistently focuses on specific segments of a time-series sequence on a per-label basis. This will enable human observers to understand the RNN's decision-making process, distinguishing data sequences into classes through attention. Attention signatures help demystify the black box, providing another tool to fine-tune, interpret, and validate time-series models.

2 Background

2.1 Related Work

Recently, progress has been made toward comprehending the internal mechanisms of neural networks, however, neural networks predominantly remain "black boxes." For text and image processing, interpretability research involves presenting the text and images in an intuitive format which can be easily verified by human observers [1,10]. In the domain of deep learning time-series models, much of the interpretability research involves analyzing the network's internal structures (e.g., neural activations), and visualizing and performing statistical analyses of input/output [5,16]. Work has also employed visualizations of RNN-CNN hybrids, where the feature maps of the CNN layers can be visualized for

interpretability [19]. Similarly, attention has been utilized with RNNs to enhance interpretability, but these efforts have primarily been in the area of feature interpretation [20]. The contribution of the work is in time-step importance interpretability per-label, providing insight into the influence of time on time-series model accuracy, giving both model builders and users further confidence in model predictions.

2.2 Recurrent Neural Networks

RNNs excel in preserving temporal information through the recurrence mechanism, which feeds the previous recurrent layer's output back into the current input. The mechanism constructs a temporal chain of causality, creating a network "memory." The memory enables the RNN to model time-series data more effectively than most other networks. Although RNNs encompass various network architectures, in the study we focus on the LSTM, which has demonstrated itself to be one of the most robust RNN variants [4,8].

Previous work has shown that attention mechanisms can significantly enhance RNN accuracy in various time-series modeling [9,14,20]. The improvement in time-series modeling arises from a RNN's inherent tendency to assign more weight to nearer time-steps than older ones. While the smooth weighting in time may be desirable in some cases, it may not be in others. Attention mitigates the issue and allows the RNN to construct possibly better representations of sequences by enabling the network, if advantageous, to attend to data "out of order" [3,9,11,18].

2.3 Attention Mechanisms

Attention mechanisms, originally created for text prediction in sequence-to-sequence models, allow the network to focus on particular portions of a sequence, highlighting the importance of one token or another within the sequence to create a more accurate representation of the sequence for output [3,11]. Attention can be broadly described as a weight or context vector of importance within a sequence [3,17].

2.4 General Attention Mechanism

The general attention mechanism creates a weight vector, known as a context vector, which captures the importance of each output step of the RNN to the prediction outcome. With access to the hidden states of the entire input sequence, the attention mechanism selects specific elements from the sequence to improve the output. In this way, the context vector enables the model to concentrate more on the relevant portions of the input sequence, as needed. See Formulas 1 through 3 [3,11].

2.5 Self-Attention Mechanism

Unlike general attention, self-attention directly calculates the importance of sequence portions relative to other portions, generating a context matrix. The context matrix enables the computation of a better representation of each sequence, allowing subsequent layers to model more effectively. Formulas 4 through 6 provide additional details on the self-attention mechanism used in the work [17,18].

3 Temporal Attention

Temporal attention in the work refers to the use of both of the aforementioned attention mechanisms in conjunction with a LSTM network in time-series modeling. We employ both self-attention and general attention (in the form of global-soft attention) to develop models that achieve both high accuracy and high interpretability. Self-attention, when added to a LSTM network, enhances accuracy by establishing relationships between different time steps [9], while general attention both further enhances accuracy (see Section 5.2) and also allows us to produce interpretable results through novel temporal attention signatures.

Temporal general attention operates by capturing the full LSTM output (i.e., "hidden layer"), typically denoted H_t^T, from within a sequence and training a separate layer to "attend" to some parts of the LSTM output more than others. Equations 1 through 3 define temporal general attention [3,11], where H is the output of the LSTM layer (i.e., input to the attention layer), and x is model input ((i.e., input to the LSTM layer).

Temporal General Attention Equations

$$e_i = \tanh(W_a H^T + b_a) \quad [\text{ similarity score }] \tag{1}$$

$$a_i = softmax(e_i) \quad [\text{ attention weights }] \tag{2}$$

$$y = \Sigma_i a_i H^T \quad [\text{ output vector }] \tag{3}$$

Temporal self-attention closely follows traditional self-attention [17,18], with minor but notable changes. The "value matrix," a projection of input, is used as-is. In the case of quantitative time-series data, there is no purpose to a transformation of the input data. Another difference is dot-product attention is used, rather than scaled dot-product. The purpose of scaling is to avoid the vanishing gradient problem. Since LSTM networks avoid this inherently, there is no purpose to scaling in the attention layer. Equations 4 through 6 describe temporal self-attention.

Temporal Self-Attention Equations

$$E = \tanh(W_a H^T + b_a) \quad [\text{ similarity score }] \tag{4}$$

$$A = softmax(W_e E) \quad [\text{ attention weights }] \tag{5}$$

$$Y = AH \quad [\text{ output matrix }] \tag{6}$$

4 Data

The data sets used are as follows and chosen with the following rationale; time-series prediction/classification tasks; comparison to similar studies [9]; sufficiently different classification tasks; a mix of data sizes and sequence lengths; both binary and multi-label classification.

- *ECG Heartbeat Categorization* (multi-label classification): ECG readings[1],[2]. Sequence size 187.
- *SmartFall* (binary classification): Raw (x, y, z) accelerometer readings of activities of daily life (ADL), predicting falls [12]. Sequence size 40.
- *Air Quality Time-Series data UCI* (binary prediction): Detecting CO levels will rise/fall tomorrow [15]. Sequence size 8.
- *Dissolved Oxygen Levels* (multi-label classification): Classifying levels of dissolved oxygen in natural bodies of water [6]. Sequence size 6.
- *Australian BOM Observations* (binary prediction): Prediction of rain/no-rain tomorrow [2]. Sequence size 5.

Larger data sets were split into train, validation, and test 60/20/20. Smaller sets are split into train and test 80/20 and run with 5-fold cross-validation.

5 Model and Methodology

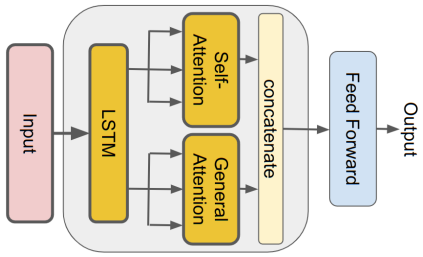


Fig. 2: Temporal attention architecture.

5.1 Architecture

The temporal attention model[3] is designed to model sequences through time while attending to portions of a time sequence in which some time-steps are more important to one another (self-attention) or to output (general attention). The combination of attention mechanisms allows for both higher accuracy (see

[1] ECG source 1.
[2] ECG source 2.
[3] Source code of the project.

Section 5.2) and uniquely interpretable results through the use of temporal attention signatures. The architecture is illustrated in Figure 2.

The use of self-attention is motivated by the desire to achieve maximal accuracy [9]. However, general attention, when used in conjunction with self-attention, further improves accuracy (see Section 5.2). Furthermore, general attention is used to generate attention signatures, which is central to the work. The choice of general attention for signature generation is based on the theoretical behavior of each type of attention. Self-attention provides information about the relative importance of different time-steps with respect to one another, effectively adding new "features" to the input and improving sequence representation. In contrast, general attention calculates absolute attention scores between input and output. Since our goal in signature generation is to determine the absolute importance of each time-step in the sequence relating to output, thereby yielding interpretable insight to the model's choices, temporal attention signatures are generated using the general attention mechanism. Self-attention was also investigated for interpretability, but did not yield per-label interpretable insights. The investigation is omitted for brevity, as it did not contribute to the study other than to further validate the theoretical choice of general attention signatures.

Table 1: Ablative study, showing temporal attention (TempAtt), removing general attention (SelfAtt), removing self-attention (GenAtt), removing both (LSTM).

Data	TempAtt	SelfAtt	GenAtt	LSTM
ECG	**0.990**	0.990	0.962	0.941
SmartFall	0.960	**0.961**	0.956	0.939
Air Quality	**0.987**	0.962	0.924	0.912
Diss Oxy	**0.976**	0.970	0.945	0.937
AUS BOM	**0.939**	0.923	0.894	0.853

5.2 Accuracy and Ablative Study

Prior work has shown that LSTMs with self-attention can achieve higher accuracy and more robust time-series models than either LSTMs alone, or self-attention alone [9]. Temporal attention as presented here uses a combination of both self-attention and general attention to further enhance accuracy. Presented in Table 1 is an ablative study demonstrating the enhanced accuracy of temporal attention over either type of single attention RNN combination and no attention (i.e., LSTM alone).

5.3 Signature Generation

Rank Matrix: The first step in signature generation is to present test sequences to a fully trained model, and to generate an attention rank matrix, which is then used to generate a confidence measure. For illustration, we use the shortest sequence length of 5 (Australian BOM data) to step through the generation and interpretation of temporal attention signatures. For each test sequence presented,

Table 2: Example rank matrix, label 0, Australian BOM weather prediction.

step	low	med-low	med	med-high	high	total	confidence
0	23	105	97	**6732**	87	7044	0.96
1	**6450**	139	172	159	124	7044	0.92
2	201	173	**6302**	146	222	7044	0.89
3	108	**6589**	203	92	52	7044	0.94
4	47	71	13	11	**6902**	7044	0.98

attention values corresponding to Equation 2 are recorded (a vector of length sequence size). The vector is normalized using min-max scaling in the range $[0, 1]$, on a *per-sequence* basis. Normalization is performed on a per-sequence basis to facilitate relative importance assessment within the sequence and for comparison between different sequences. Normalized scores are ranked in order of importance (lowest importance 0 to highest important 4). The rankings of all normalized attention vectors per-label are compiled into a rank matrix. The matrix represents the per-label frequency in which a time-step per sequence occurs at each importance ranking (see Table 2). Reading the matrix row-wise determines the per-label frequency of each time-step at each importance ranking. For example, in Table 2, the rank matrix for label 0 of the Australian BOM data shows time-step 0 is of medium-to-high importance most of the time (6,732/7,044 or 96%), while time-step 1 is of low importance most of the time (6,450/7,044 or 90%), and so on.

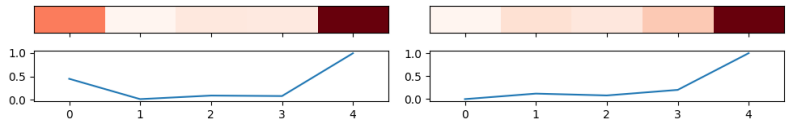


Fig. 3: Sample signatures (labels [0,1]) for the Australian BOM rain prediction.

Confidence Measure: Using the rank matrix, a confidence vector is calculated as shown in Table 2's *confidence* column. If the average of the vector is 90% or higher (determined empirically), the time-steps' rankings are considered "confident." For example, in Table 2, the confidence is 0.94. Once judged confident, the normalized attention scores on a per-label basis are averaged to calculate a single normalized attention vector per-label. These single vectors represent each label's *attention signature*. Figure 3 displays the attention signatures for the "no rain" (left) and "rain" (right) prediction labels for the Australian BOM weather data.

6 Experiments and Results

Each data set was trained repeatedly until 20 models with both high accuracy (> 0.92) and confidence (> 0.90) were obtained. All models were processed as described in Section 5.3 and the stability and repeatability of the signatures were verified for each data set (see next paragraph), producing a set of per-label temporal attention signatures per data set. Each of these signature sets is then be analyzed for interpretability insights (see Section 7.1).

To validate that attention signatures are similar across models, we calculate the Euclidean distances between per-label signatures per-model. Since signature values for each time step are scaled to $[0, 1]$, the maximum possible Euclidean distances between signatures is known. Therefore, along with visual inspection of the graphical signatures, average Euclidean distances between models may be used as a validation measure to ensure signatures are stable and repeatable. Table 3 shows these averages and maximums across experiments and demonstrates that signatures across models are similar, stable, and repeatable. Figures 3 through 7 detail the temporal signatures of each of the data sets.

Table 3: Euclidean distances summaries for all data sets.

data set	average e-dist	max e-dist	sequence length	percent max
Australian BOM	0.128	2.24	5	05.7%
Air Quality	0.581	4.47	20	13.0%
SmartFall	0.771	6.32	40	12.2%
Dissolved O2	0.185	2.45	6	07.5%
ECG	1.631	10.0	187	16.3%

7 Applications

7.1 Interpretability

Visualizations of temporal attention signatures enable the identification of important time intervals within a sequence, providing insight into the most relevant factors for classification. This helps both validate the model and also enhances our understanding of critical data points that may require attention when making human decisions. For instance, the Australian BOM weather signatures (see Figure 3) clearly demonstrate that step 5 is of high importance to both labels. This is intuitive, since yesterday's weather is a major factor in predicting today's weather. However, the "no rain" label signature also indicates that weather five days ago carries some importance. This provides insight that, further into the past, there may be information that could aid in the prediction of one case versus another.

To further analyze our findings, we select two more complex data sets, the SmartFall ADL data (multivariate, binary label) and ECG data (univariate, multi-label), for more in-depth discussion (see Figures 6 and 7 respectively). In

each case, a typical input sequence for each label is overlaid with the attention signature (dotted blue line). In the SmartFall data set, the orange, green, and red lines represent the x, y, and z accelerometer scores. In the ECG data set, the single-channel ECG reading is in orange.

In the case of the SmartFall data, label 1 (fall) exhibits the network attending to data immediately preceding the fall and the fall itself (steps 20-35). In contrast, for label 0 (no fall), the network attends to the majority of the input data (steps 5-40). Both results are intuitive and demonstrate that the network attends to a fall much in the same way a human would. When observing a non-fall ADL, a human would tend to observe most things equally. However, when a fall is observed, human attention immediately narrows to the fall and the time immediately preceding the fall to detect "what happened?" This clearly indicates that the network can distinguish between "fall" and "no fall" similarly to human observation.

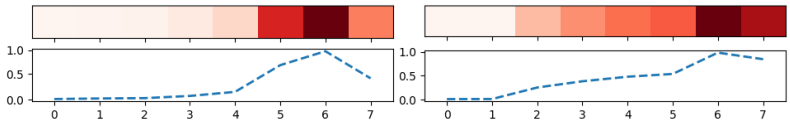


Fig. 4: Air quality data. Label 0: CO levels fall (left), and label 1: CO levels rise (right).

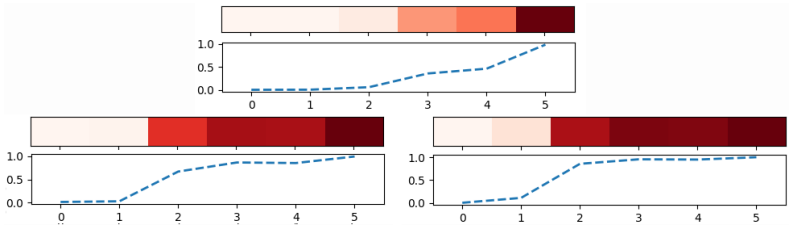


Fig. 5: Dissolved oxygen data. Label 0: toxic (top), label 1: inhospitable (bottom-left), label 2: supports life (bottom-right).

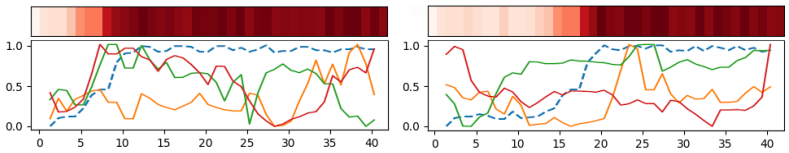


Fig. 6: SmartFall data. Label 0: no fall (left), and label 1: fall (right). Line graphs show temporal attention signature (dotted blue) and x,y,z signals (orange, green, red).

In the case of ECG data, for label 1, which corresponds to superventricular ectopic beats (i.e., premature and/or narrow beats), the network attends to the early portion of the cycle as expected. For label 2, which corresponds to ventricular ectopic beats (i.e., small changes in normal heartbeats leading to

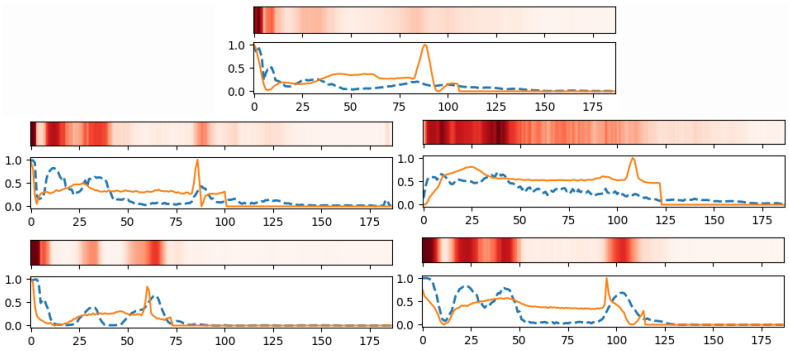


Fig. 7: ECG data. Label 0: Non-Ectopic (top), label 1: Superventrical Ectopic (mid-left), label 2: Ventricular Ectopic (mid-right), label 3: Fusion (btm-left), label 4: Unknown (btm-right). Lines show temporal attention (dotted blue) with ECG signal.

extra or skipped beats), the network pays more general attention throughout the cycle due to its complex nature. For label 3, which corresponds to fusion beats (i.e., supraventricular and ventricular impulses coincide), the network attends to the fusion portions as expected. For label 0, which corresponds to normal beats, the network attention is low throughout most of the cycle, indicating nothing noteworthy. Lastly, for label 4, which corresponds to unknown beats, we observe temporal attention attending with a strong correlation to the beat cycle itself, indicating that the shape of the beat is noteworthy and may require further investigation by a domain expert.

These observations provide two valuable insights. First, we can validate our intuition by correlating network attention with domain expert knowledge, thereby increasing confidence in the model. Secondly, we can use temporal attention to gain new insight into unknown states by investigating data portions that the networks indicate are important.

Not discussed for brevity, the two remaining signature sets, dissolved oxygen and air quality data, show similar results; clearly distinguishable temporal attention signatures which domain experts may examine for validation and insight.

7.2 Sequence Length

In the context of time-series modeling, selecting optimal sequence length is not always straightforward, especially when the data does not possess a natural fixed sequence length. Typically, sequence length selection is based on intuition, trial-and-error, or grid search. By generating temporal attention signatures, a relationship between sequence length and prediction becomes clearly evident, which provides insight into both selecting and empirically validating sequence length. Based on the observations made by temporal attention signatures thus far, it should be expected that selecting a sequence length that is too long will yield a signature with very little importance toward the beginning of the sequence. Similarly, selecting a sequence length that is too short will yield a signature with

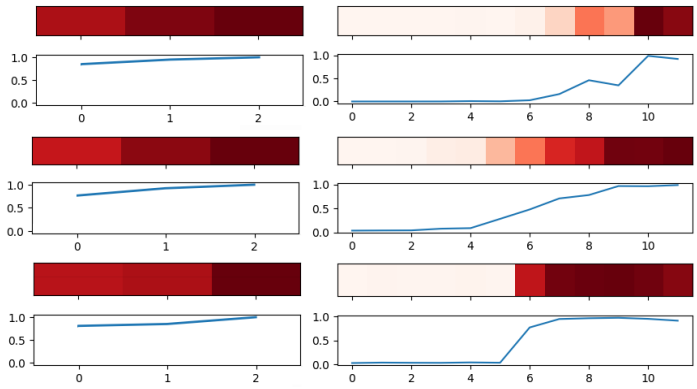


Fig. 8: Dissolved oxygen data set signatures (labels 0, 1, 2) for sequence size 3 (left) and sequence size 12 (right).

most steps showing high importance. Therefore, temporal attention signatures can be used to reduce the guesswork and cost associated with obtaining the optimal sequence length, as well as to validate the final sequence length selection, regardless of how it was chosen.

As empirical evidence, we examine the dissolved oxygen data set, where an optimal sequence size of 6 was obtained through grid search. Figure 8 illustrates training using size 3 (left) and size 12 (right). When the sequence size is low (3), all time steps are heavily weighted, indicating that there is more relevant information further into the past. At sequence size 12, almost all weighting is assigned to the six most recent time steps. Furthermore, the model accuracies are as follows: size 6: 0.98, size 3: 0.91, size 12: 0.95.

For brevity, demonstration with other data sets is omitted, however extensive sequence length experimentation validates the result. This demonstrates that visual inspection of temporal attention signatures can be employed to quickly narrow the field of sequence size choices, and can also be used to validate sequence length, regardless of the selection method.

8 Conclusion

The work presents a novel and practical approach for simultaneously increasing accuracy in time-series modeling and analyzing time-series data through temporal attention. Temporal attention signatures allow for the identification of important events or patterns within the data that contribute to the final prediction or classification. By analyzing attention signatures, one can validate or challenge prior assumptions, providing confidence in the model. Additionally, the technique can reveal new insights previously unknown to the analyst. Furthermore, attention signatures can aid in the determination of optimal sequence length, reducing the need for costly trial-and-error methods. Ultimately, the approach inherently improves accuracy and also leads to further improvements by facilitating the fine-tuning of sequence length through visual inspection.

References

1. An, J., Joe, I.: Attention map-guided visual explanations for deep neural networks. Applied Sciences **12**, 3846 (04 2022)
2. Australian Bureau of Meteorology (BOM): Australia, Rain Tomorrow. Australian BOM National Weather Observations
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. ArXiv **1409** (09 2014)
4. Cheng, J., Dong, L., Lapata, M.: Long short-term memory-networks for machine reading. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 551–561 (01 2016)
5. Davel, M., Theunissen, M., Pretorius, A., Barnard, E.: Dnns as layers of cooperating classifiers. Proceedings of the AAAI Conference on Artificial Intelligence **34**, 3725–3732 (04 2020)
6. Durell, L., Scott, J.T., Hering, A.S.: Replication Data for: Functional Forecasting of Dissolved Oxygen in High-Frequency Vertical Lake Profiles (2022)
7. Guo, H., Fan, X., Wang, S.: Human attribute recognition by refining attention heat map. Pattern Recognition Letters **94** (05 2017)
8. Hewamalage, H.: Recurrent neural networks for time series forecasting: Current status and future directions. International Journal of Forecasting **37** (08 2020). https://doi.org/10.1016/j.ijforecast.2020.06.008
9. Katrompas, A., Ntakouris, T., Metsis, V.: Recurrence and self-attention vs the transformer for time-series classification: A comparative study. In: Artificial Intelligence in Medicine. pp. 99–109. Springer International Publishing (2022)
10. Liang, Y., Li, M., Jiang, C.: Generating self-attention activation maps for visual interpretations of convolutional neural networks. Neurocomputing **490** (11 2021)
11. Luong, M.T., Pham, H., Manning, C.: Effective approaches to attention-based neural machine translation (08 2015)
12. Mauldin, T., Canby, M., Metsis, V., Ngu, A., Rivera, C.: Smartfall: A smartwatch-based fall detection system using deep learning. Sensors **18**, 3363 (10 2018)
13. McClarren, R.: Recurrent Neural Networks for Time Series Data, pp. 175–193 (05 2021)
14. Qin, Y., Song, D., Cheng, H., Cheng, W., Jiang, G., Cottrell, G.: A dual-stage attention-based recurrent neural network for time series prediction (04 2017)
15. S. De Vito et al.: On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario
16. Siddiqui, S., Mercier, D., Munir, M., Dengel, A., Ahmed, S.: Tsviz: Demystification of deep learning models for time-series analysis. IEEE Access **PP**, 1–1 (04 2019)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention is all you need. 31st Conference on Neural Information Processing Systems (NIPS 2017) (06 2017)
18. Vaswani, P., Uszkoreit, J., Shaw, A.: Self-attention with relative position representations. pp. 464–468 (01 2018)
19. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification. 2016 IEEE Conference on Computer Vision and Pattern Recognition (04 2016)
20. Zhang, X., Liang, X., Li, A., Zhang, S., Xu, R., Wu, B.: At-lstm: An attention-based lstm model for financial time series prediction. IOP Conference Series: Materials Science and Engineering **569**, 052037 (08 2019)