

# P2ExNet: Patch-based Prototype Explanation Network

Dominique Mercier<sup>1,2</sup>✉, Andreas Dengel<sup>1,2</sup>, and Sheraz Ahmed<sup>2</sup>

<sup>1</sup> Computer Science Department, Technische Universität Kaiserslautern,  
67663 Kaiserslautern, Germany

<sup>2</sup> Smart Data and Knowledge Services, German Research Center for Artificial  
Intelligence (DFKI), 67663 Kaiserslautern, Germany  
{dominique.mercier, andreas.dengel, sheraz.ahmed}@dfki.de

**Abstract.** Deep learning methods have shown great success in several domains as they process a large amount of data efficiently, capable of solving difficult classification, forecast, segmentation, and other tasks. However, these networks suffer from their inexplicability that limits their applicability and trustworthiness. Although there exists work addressing this perspective, most of the existing approaches are limited to the image modality due to the intuitive and prominent concepts. Unfortunately, the patterns in the time-series domain are more complex and non-comprehensive, and an explanation for the network decision is pivotal in critical areas like medical, financial, or industry. Addressing the need for an explainable approach, we propose a novel interpretable network scheme, designed to inherently use an explicable reasoning process inspired by the human cognition without the need of additional post-hoc explainability methods. Therefore, the approach uses class-specific patches as they cover local patterns, relevant to the classification, to reveal similarities with samples of the same class. Besides, we introduce a novel loss concerning interpretability and accuracy that constraints P2ExNet to provide viable explanations of the data that include relevant patches, their position, class similarities, and comparison methods without compromising performance. An analysis of the results on eight publicly available time-series datasets reveals that P2ExNet reaches similar performance when compared to its counterparts while inherently providing understandable and traceable decisions.

**Keywords:** Deep Learning · Convolutional Neural Networks · Time-Series Analysis · Data Analysis · Explainability · Interpretability.

## 1 Introduction

Nowadays, deep neural networks are popular and used in many different domains comprising image processing, natural language processing, and time-series processing. Though these deep networks have achieved high performance, they are still black boxes in nature. This behavior makes it tough to understand the reasons behind the decisions. In particular, this black box nature hinders the use

of these models in critical domains like medical, autonomous driving, industrial, financial, and raises the need for interpretability methods to provide intuitive and understandable explanations. Only explainable models can be usable in critical domains that require transparency [16].

The existing methods for interpreting decisions of deep learning models are mostly applicable to image modalities. In particular, image concepts are intuitive by default [25]. Besides the image domain, there is only a limited amount of work in the field of time-series as the modalities are more complex and usually not directly interpretable for a human. Nevertheless, these time-series analysis networks and their explanations are pivotal for their industrial and financial use. Therefore, we propose P2ExNet as an approach to deal with time-series data.

Also, existing approaches are mostly post-hoc methods that are applied after the classifier to explain their decisions [7]. Intuitively, these approaches keep the network as it is without any change to the structure, enabling their use on almost every architecture. Usually, this results in an instance-based local explanation that does not explain any global behavior. In contrast to post-hoc methods, the intrinsic methods focus on model design concerning the inference process to provide an understandable global explanation. Ultimately, neither of the two approaches is superior as both have to deal with several limitations regarding the quality, subjectivity [14], the audience, and the domain usage.

To overcome these limitations, we propose a network architecture for time-series analysis based on the standard deep neural network architecture providing a global explanation using representative class-specific prototypes and an instance-based local explanation using patch-based similarities and class-similarities. The inference process of our architecture follows the human-related reasoning process [11] and uses concepts and prototypes [13]. Intuitive class-specific patches explain the network decision. Our approach is superior compared to existing template matching approaches [5] in the manner of generalization and applicability. Our experiments emphasize the use of our network structure by highlighting the comparable performance when compared to a non-interpretable network of the same size over eight publicly available datasets while preserving an intuitive and traceable explanation.

## 2 Related Work

The field of network interpretability covers post-hoc and intrinsic methods. Based on the use-case, it is not always possible to use both methods as these methods come with restrictions concerning the data and the network. In the following paragraphs, we address the perspectives, their advantages, and drawbacks.

### 2.1 Post-hoc

Using post-hoc methods to explain the decisions of deep neural networks is a very prominent approach as these methods do not modify the network architecture and can provide an instance based explanation. Furthermore, these methods offer instance-based as well as global explanations resulting in broad applicability.

**Instance-based:** A widespread instance-based post-hoc class of approaches in the field of image domain are so-called back-propagation methods [4]. These approaches produce heat-maps highlighting the most relevant and sensitive parts concerning the network decision. There exist enhancements that evolved [27] and take various aspects into account to improve the expressiveness and consistency. Another post-hoc instance-based class of methods are the layer-wise relevance propagation methods [3, 10] that produce results that are close to the heat-maps but more stable. In particular, the image domain explored different approaches to visualize the activations [24] or make use of the gradients [18] or saliency [21] to produce heat-maps for instances. However, in the case of the time-series modalities, there exists only a limited amount of work [20].

**Global:** In contrast to instance-based methods, there exist attempts to compute a global behavior based on the influence of the samples [12, 23]. These methods provide an idea of helpful and harmful dataset samples to detect outliers and debug dataset using the sample influence. Another approach is to attach an interpretable architecture to the trained network. As presented in [15], the attachment of an autoencoder before the neural network and a customized loss function for the autoencoder can enhance the interpretability. Siddiqui et al. [19] presented an adoption of this approach for the time-series domain with an adjusted loss function.

## 2.2 Intrinsic

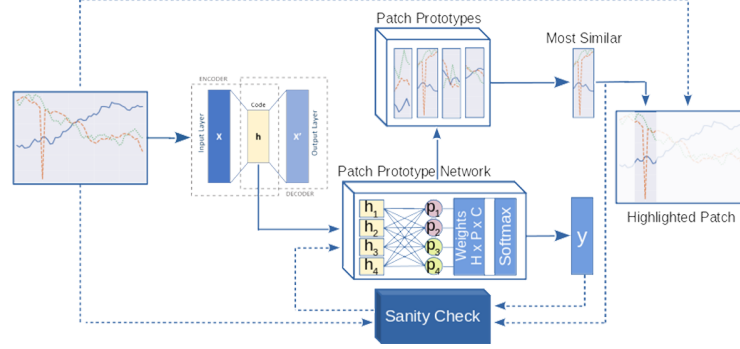
Intrinsic methods approach the problem from a different perspective by incorporating the interpretability directly. Therefore, they modify the model architecture by introducing interpretable layers [26]. A drawback of these approaches is the restricted learning process that can harm the performance. An intuitive interpretable layer solution are prototype layers to explain model decision [2]. Mainly, two types of prototypes showed to provide reasonable explanations. First, class prototypes that cover the complete input [8, 13] and second patch prototypes [6].

## 2.3 Limitations of Existing Methods

Even though there exists work to explain the network decisions, most of the approaches are limited to image modalities [17]. Furthermore, there is ongoing research investigating the consistency, expressiveness, and subjectivity of these explanations. Some findings prove the inconsistency of saliency-based methods [22] and the expressiveness [1]. Also, methods that use sparsity constraints suffer from the same problems concerning their consistency.

## 3 P2ExNet: The Proposed Approach.

This section provides insights into the proposed approach. It starts with a motivation followed by the general architecture structure, the mathematical background, and the training procedure.



**Fig. 1. Inference and testing workflow.** Artificially, computed prototypes are evaluated in a similarity-based manner to suggest class-specific patches.

### 3.1 Motivation: An Understandable Reasoning Behavior.

Inspired by human reasoning behavior, we aligned our framework to rely on implicit knowledge about objects and examples already seen before. This approach is similar to the humans’ inference process. Precisely, we compare new instances to abstract concepts include class-specific features. The term prototypical knowledge describes the knowledge about these concepts and covers the analogical process to map new to the existing knowledge [9]. Following this process, the proposed method uses shallow representations. These prototypes encode class-specific pattern and provide the decision based on similarity.

### 3.2 Architecture

Inspired by the work of Gee et al. [8], we combined an autoencoder with a prototype network. The autoencoder consists of several convolutional and max-pooling layers serving as a feature encoding network to provide a latent representation that encodes the relevant features of an input sequence. This representation is fed forward to a custom prototype layer to generate prototypes. Motivated by the work of Chen et al. [6], we use multiple prototypes to represent a sample rather than a single one for the complete input. Precisely, the prototype layer has randomly initialized variables representing patch prototypes of user-defined size. Larger sizes will result in composed concepts, and smaller sizes result in more basic concepts. On top of the prototype layer, we attached a prototype-weight layer to encourage class-specific prototypes and weight their position within the sample to cover the local importance. Finally, a soft-max classification evaluates the similarity scores produced by the prototype layer multiplied with weights of the prototypes, as shown in Figure 1.

### 3.3 Mathematical Background

Our method uses a novel combined loss that captures several aspects enabling the network to produce a meaningful set of patch prototypes based on the losses

proposed by [6, 8]. For the following equations, let  $S_x$  be the set of patches corresponding to a sample  $x$  and the set  $P$  of prototypes.

**Distances:** We use the  $L^2$  norm to compute the distance between any two vectors. Furthermore, we compute the minimum distance between a sample and any prototype ( $D_{s2p}$ ) and vice versa ( $D_{p2s}$ ). We denote  $D_{p2p}$  as the minimal distance between a prototype and all others and calculate the minimum distance to a prototype of the same class  $D_{clst}$  and to the other classes  $D_{sep}$  w.r.t.  $y$ . Therefore  $P_y$  denotes the subset of  $P$  assigned to the class label of  $y$ . The distances are shown in Equations 1 to 4.

$$D_{s2p}(s) = \min_{p \in P} L^2(s, p) \quad (1) \quad D_{p2p}(p) = \min_{p' \in P} L^2(p, p') \quad (2)$$

$$D_{clst}(s, y) = \min_{p \in P_y} L^2(s, p) \quad (3) \quad D_{sep}(s, y) = \min_{p \in \{P \setminus P_y\}} D_{s2p}(s, p) \quad (4)$$

**Losses:** To ensure high-quality prototypes, we introduce our novel patch loss. This loss is a combination of different objectives to achieve good accuracy and an explanation that does not contain duplicates or prototypes that are not class-specific. Our loss combines the following losses:

- **Autoencoder loss:** MSE is used to encourage reconstruction later used for prototype reconstruction.
- **Classification loss:** To produce logits for the softmax cross-entropy we multiply the reciprocal of  $D_{s2p}$  and the prototype-weight layer.
- $L_{p2s}$  **and**  $L_{s2p}$ : These losses preserve the relation between the input and the prototypes and vice versa as shown in Equation 5 and 6.
- $L_{div}$ : The diversity among the patch prototypes is computed as shown in Equation 7.
- $L_{clst}$  **and**  $L_{sep}$ : To encourage the network to learn class-specific prototypes we compute  $L_{clst}$  and similarly to  $L_{sep}$  but with a negative sign. This penalized prototypes close to samples of the wrong class w.r.t. their assigned class.

$$L_{p2s}(x) = \frac{1}{|P|} \sum_{p \in P} D_{p2s}(p) \quad (5) \quad L_{s2p}(x) = \frac{1}{|S_x|} \sum_{s \in S_x} D_{s2p}(s) \quad (6)$$

$$L_{div} = \log(1 + \frac{1}{|P|} \sum_{p \in P} D_{p2p}(p))^{-1} \quad (7) \quad L_{clst}(x, y) = \frac{1}{|S_x|} \sum_{s \in S_x} D_{clst}(s, y) \quad (8)$$

Our proposed final loss is a linear combination taking into account previously mentioned aspects and ensures meaningful, diverse, and class-specific patch prototypes shown in Equation 9. By default, we set all lambda values except  $\lambda_c$  to one to find the best compromise between the objectives preserving high accuracy.

$$\begin{aligned} Patch\_Loss(x, y) = & \lambda_c H(x, y) + \lambda_{mse} MSE(x, x) + \lambda_{p2s} L_{p2s}(x) \\ & + \lambda_{s2p} L_{s2p}(x) + \lambda_{div} L_{div} + \lambda_{clst} L_{clst}(x, y) + \lambda_{sep} L_{sep}(x, y) \end{aligned} \quad (9)$$

### 3.4 Training Process

The training process of your approach consists of two stages. In the first stage, we fix the weights of the pre-initialized prototype-weight layer to ensure class-specific prototypes. We then train the network until it converges. In the second learning phase, all layers except the prototype-weighting layer are frozen, and the network learns to adjust the prototype weights. The adjustment corrects the prototype class affiliation using the previously trained latent representation.

## 4 Datasets

We used eight publicly available time-series datasets to emphasize the broad applicability of our approach and examine possible limitations. As a representative set, we used seven different datasets from the UCR Time Series Classification Repository<sup>1</sup> and a point anomaly dataset proposed in [20]. These datasets and their parameters are visualized in Table 1. Note that the Devices dataset corresponds to the 'Electrical devices' dataset taken from the UCR. To have better coverage of different types, we selected the datasets based on the characteristics concerning the number of classes, channels, and time-steps to cover several conditions and show the prototypes. However, we focus on classification datasets.

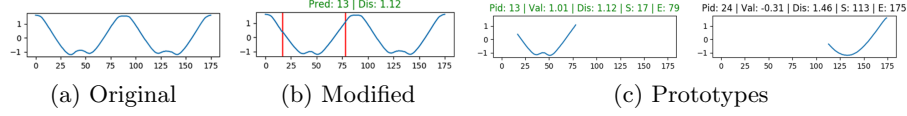
## 5 Experiments

In this section, we present our results concerning the performance, applicability, and resource consumption for our proposed approach, highlighting a comparable performance while producing interpretable results.

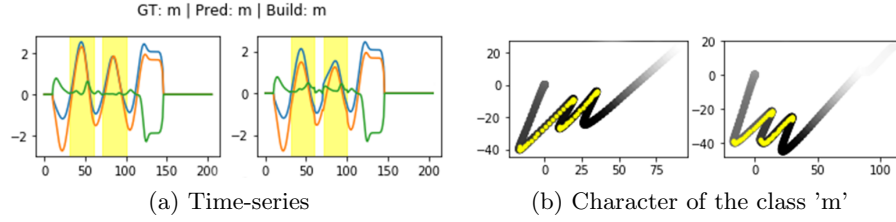
### 5.1 P2ExNet: Instance-based Evaluation

The proposed method provides the possibility to identify and highlight the parts of the input that were most relevant for the classification. Besides, it provides prototypes along with a sample containing the prototypes to compare it to the original input. Figure 2 shows highlighted regions that were important for the inference on the ADIAC dataset sample. This explanation includes the original sample of the adiac dataset, a modified version, and two prototypes. In the modified version shown in Figure 2b, we replaced the part between the two red lines with the most important patch prototype to show how close it is to the original part. Figure 2c shows two prototypes. The value of each prototype denoted as 'Val' highlights its contribution towards the classification result. Similarly, Figure 3 shows a sample from the character trajectories dataset and the mapping of the time-series back to the character. The black value highlights the pressure of the pen, and the yellow part shows the mapping of the prototype back to the input space. In the case of an incorrect classification, the prototypes

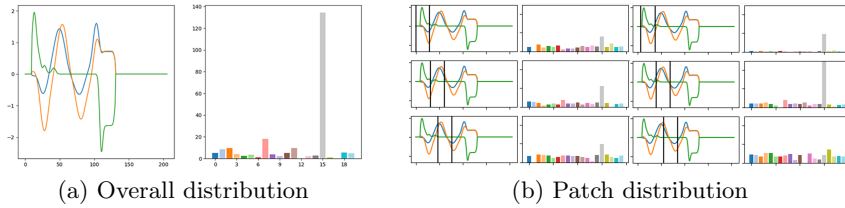
<sup>1</sup> <http://www.timeseriesclassification.com/>



**Fig. 2. Adiac dataset prototype explanation.** a) shows the original series. b) shows the series with the prototype between the red bars. c) shows two prototypes.

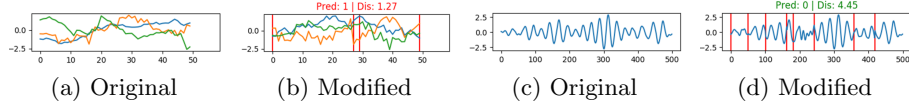


**Fig. 3. Character dataset prototype explanation.** a) shows the original series and the series with the prototypes. b) shows the character output and the modified character.



**Fig. 4. Class and prototype distribution.** a) shows the class similarities. b) shows some patches and the corresponding class similarities.

have a red caption. Furthermore, in Figure 4 the class-wise overall and patch-wise distribution provides additional information about similar classes and important patch positions. Especially in Figure 4b, we show that not all patches have the same importance when it comes to the classification. There are sensitive datasets for which the re-classification can change if the original data gets replaced with a prototype. However, for the classification and the explanation, this is not a problem as it can be solved. A proper re-scaling and adjustment can remove the offset between the prototype and the time-series. In Figure 5b such a jump in the orange signal is shown and leads to an anomaly. However, the classification of the original signal with the network was correct. Furthermore, some datasets are invariant to small offsets shown in Figure 5d. That is why re-scaling should be done based on the problem task. In case of a point anomaly task, the patches have to align. In a classification task, it is unlikely that the offset of a single point changes the prediction.



**Fig. 5. Prototype substitution.** a) and c) show original time-series. b) and d) show the corresponding modified samples and their re-classification.

**Table 1. Accuracy comparison.** A comparison of interpretable and the corresponding non-interpretable counterpart.

Dataset	Classes	Length	Channel	CNN	P2ExNet
Anomaly [20]	2	50	3	<b>99.79</b>	93.79
FordA	2	500	1	85.44	<b>89.32</b>
Devices	7	96	1	55.42	<b>62.53</b>
Adiac	37	176	1	<b>63.54</b>	60.15
Crop	24	46	1	68.27	<b>68.54</b>
50words	13	270	1	76.84	<b>81.98</b>
PenDigits	10	8	2	<b>94.29</b>	93.95
Character	20	206	3	<b>96.53</b>	91.78

## 5.2 P2ExNet: Evaluation as a Classifier

Usually, intrinsic interpretability approaches come with an accuracy drop. In Table 1 we present the accuracy trade-off highlighting that our structure is on the same level as the non-interpretable counterpart. To create a network similar to ours without the interpretable part, we replaced the prototype layer with a dense layer and a cross-entropy loss, as suggested by Chen et al. [6]. Furthermore, we removed the decoder as there is no need to restrict the latent representation as no reconstruction is required. We conducted this comparison for all eight datasets showing that P2ExNet achieves comparable or better performance in comparison to the non-interpretable variant. Overall the interpretable network has an insignificant performance increase of 0.03%. Each architecture was superior in four out of the eight datasets. The results show that the accuracy using the interpretable model dropped about 6% on the anomaly dataset but increased 7% on the Electric Devices dataset.

## 5.3 P2ExNet: Sanity Check

To prove the class-specific and meaningful behavior of the prototypes, we replaced the original time-series once with the most positive and once with the most negative influencing prototypes. In Table 2 we show that the replacement with the most confident prototypes corresponding to the predicted class achieved results close to the default accuracy, whereas the best fit prototype of a different class dramatically decreased the performance as the prediction switched. These results show that our prototypes are class-specific. However, we conducted the



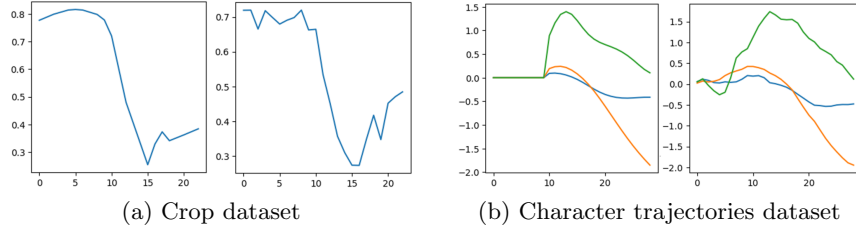
**Table 2. Replacement of original patch.** The second column shows how much data was replaced with the suggested prototypes proposed by P2ExNet. The third column shows whether the prediction was the same as with the original time-series or not. The fourth column shows the P2ExNet accuracy for the original sample and the last column for the sample replacing the original patch with the suggested patch. The first row of each dataset corresponds to replacements with the most similar whereas the second row with the most different prototype.

Dataset	Data replaced	Equal Pred.	P2ExNet Acc.	P2ExNet mod. Acc.
Anomaly	71.99	87.43	93.79	<b>91.78</b>
	67.32	19.45		22.72
FordA	51.17	99.92	89.32	<b>89.40</b>
	44.95	23.09		32.69
Devices	52.36	81.65	62.53	<b>60.52</b>
	65.81	49.81		39.11
Adiac	35.22	85.97	60.15	<b>55.98</b>
	69.90	9.11		14.84
Crop	50.50	94.08	68.54	<b>66.94</b>
	81.12	22.01		23.28
50words	36.43	93.01	81.98	<b>77.20</b>
	52.88	62.50		56.98
PenDigits	69.47	99.31	93.95	<b>93.54</b>
	68.65	8.83		11.0
Character	18.15	92.93	91.78	<b>85.30</b>
	52.90	31.71		32.87

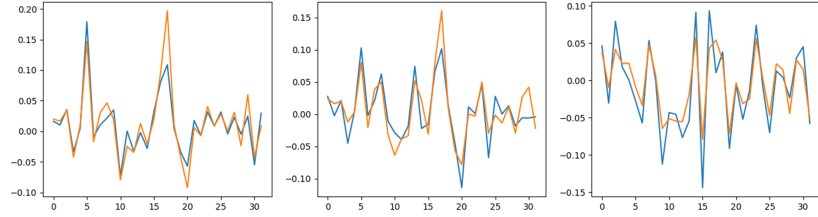
**Table 3. Closeness of prototypes.** The difference between representative and generated latent patch prototypes for P2ExNet with and without the use of the decoder are shown.

Dataset	P2ExNet with decoder	P2ExNet without decoder	Improvement
Anomaly	0.6393	<b>0.4929</b>	-22.9%
FordA	<b>0.7018</b>	1.0315	47.0%
Devices	0.4135	<b>0.3399</b>	-17.8%
Adiac	0.538	<b>0.4993</b>	-6.2%
Crop	<b>0.442</b>	0.4815	8.9%
50words	<b>0.0413</b>	0.2086	505.1%
PenDigits	<b>0.5123</b>	0.5622	9.7%
Character	<b>0.0099</b>	0.5887	5946.5%

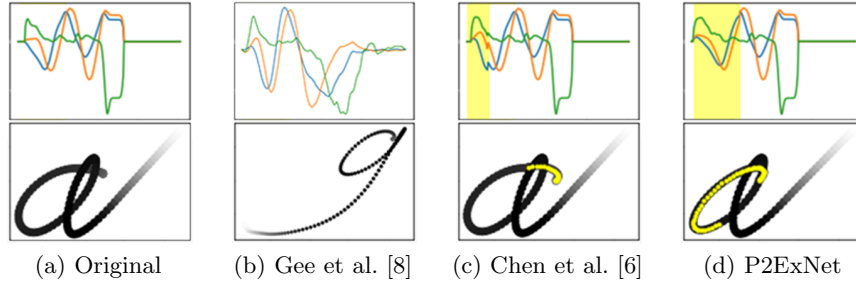
second sanity check to investigate the need for the decoder to produce latent representations that are close to the representative prototypes. In Table 3 we show that for the character trajectories, 50words, and the FordA dataset there is a significant difference if the decoder gets excluded. Also, we compared the representative and decoded prototypes and visualized two prototypes in Figure 6 highlighting the small difference between the selected representative sample (left) and the decoded one (right). We further provide the latent representation of the



**Fig. 6. Prototype comparison.** This figure shows the representative patch based on the distance to the latent prototype and the reconstruction of the latent representation.



**Fig. 7. Latent space difference.** The difference between the prototype (orange) and the real sample (blue) in the latent space for each channel are shown.



**Fig. 8. P2ExNet approaches.** Different explanations of the character 'a'.

character trajectory prototype in Figure 7. Each plot represents one of the three channels and the blue color encodes the part of the selected sample whereas the orange color decodes the latent representation of the prototype. It is clearly visible that both latent representations share the same pattern and therefore result in a similar decoded representation as shown in Figure 6b.

#### 5.4 Comparison with Existing Prototype-based Approaches

We compared the proposed method against existing work [6] and [8]. Precisely, we highlight the explanations and additional outputs. In Figure 8 we show the explanation of each approach for a character 'a' sample. While [8] explains

the class with a prototype providing a single prototype capturing the complete sample, [6] is based on parts of the input leading to a more detailed explanation. This method searches a patch for a region in the input image. Precisely, this means additional position information is available. Lastly, our proposed method provides the same information about the location but offers re-scaling as well as an implicit comparison to other prototypes and a class distribution for the complete sample and the patches, as shown in Figure 4b. Furthermore, our prototypes are class-specific and invertible. It is possible to decode them for a comparison with the representatives.

## 6 Conclusion

Summarizing our results, we came up with novel network architecture, along with a loss and training procedure aligned to produce interpretable results and an inference process similar to the human reasoning without a significant drop in performance. Further, we proved that the proposed method works for several time-series classification tasks and when excluding the class-specific prototype assignment, our approach is suitable to produce prototypes for regression and forecast tasks. Besides, we compared the proposed method with existing prototype-based methods concerning their interpretable output and time consumption, finding ours superior in both aspects.

## Acknowledgements

This work was supported by the BMBF projects DeFuseNN (Grant 01IW17002) and the ExplAINN (BMBF Grant 01IS19074). We thank all members of the Deep Learning Competence Center at the DFKI for their comments and support.

## References

1. Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049 (2018)
2. Angelov, P., Soares, E.: Towards explainable deep neural networks (xdnn). arXiv preprint arXiv:1912.02523 (2019)
3. Arras, L., Montavon, G., Müller, K.R., Samek, W.: Explaining recurrent neural network predictions in sentiment analysis. arXiv preprint arXiv:1706.07206 (2017)
4. Bojarski, M., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., Muller, U., Zieba, K.: Visualbackprop: efficient visualization of cnns. arXiv preprint arXiv:1611.05418 (2016)
5. Brunelli, R.: Template matching techniques in computer vision: theory and practice. John Wiley & Sons (2009)
6. Chen, C., Li, O., Tao, C., Barnett, A.J., Su, J., Rudin, C.: This looks like that: deep learning for interpretable image recognition. arXiv preprint arXiv:1806.10574 (2018)
7. Choo, J., Liu, S.: Visual analytics for explainable deep learning. IEEE computer graphics and applications **38**(4), 84–92 (2018)

8. Gee, A.H., Garcia-Olano, D., Ghosh, J., Paydarfar, D.: Explaining deep classification of time-series data with learned prototypes. *arXiv preprint arXiv:1904.08935* (2019)
9. Gentner, D., Colhoun, J.: Analogical processes in human thinking and learning. In: *Towards a theory of thinking*, pp. 35–48. Springer (2010)
10. Gu, J., Yang, Y., Tresp, V.: Understanding individual decisions of cnns via contrastive backpropagation. In: *Asian Conference on Computer Vision*. pp. 119–134. Springer (2018)
11. Guidoni, P.: On natural thinking. *The European Journal of Science Education* **7**(2), 133–140 (1985)
12. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. pp. 1885–1894. JMLR. org (2017)
13. Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: *Thirty-Second AAAI Conference on Artificial Intelligence* (2018)
14. Lipton, Z.C.: The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016)
15. Palacio, S., Folz, J., Hees, J., Raue, F., Borth, D., Dengel, A.: What do deep networks like to see? In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018)
16. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017)
17. Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., Keim, D.A.: Towards a rigorous evaluation of xai methods on time series. *arXiv preprint arXiv:1909.07082* (2019)
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 618–626 (2017)
19. Siddiqui, S.A., Mercier, D., Dengel, A., Ahmed, S.: Tsinsight: A local-global attribution framework for interpretability in time-series data. *arXiv preprint arXiv:2004.02958* (2020)
20. Siddiqui, S.A., Mercier, D., Munir, M., Dengel, A., Ahmed, S.: Tsviz: Demystification of deep learning models for time-series analysis. *IEEE Access* **7**, 67027–67040 (2019)
21. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013)
22. Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., Preece, A.: Sanity checks for saliency metrics. *arXiv preprint arXiv:1912.01451* (2019)
23. Yeh, C.K., Kim, J., Yen, I.E.H., Ravikumar, P.K.: Representer point selection for explaining deep neural networks. In: *Advances in Neural Information Processing Systems*. pp. 9291–9301 (2018)
24. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015)
25. Zhang, Q.s., Zhu, S.C.: Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* **19**(1), 27–39 (2018)
26. Zhang, Q., Nian Wu, Y., Zhu, S.C.: Interpretable convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8827–8836 (2018)
27. Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M.: Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595* (2017)