ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.

The Coming of Age of Interpretable and Explainable Machine Learning Models

P.J.G. Lisboa[1], S. Saralajew[2], A. Vellido[3], and T. Villmann[4] *

1 - Liverpool John Moores University, Liverpool - United Kingdom

2 - Bosch Center for Artificial Intelligence, Renningen - Germany

3 - Universitat Politècnica de Catalunya and IDEAI-UPC Research Center,
Barcelona - Spain

4 - University of Applied Sciences Mittweida,
Saxon Institute for Comp. Intelligence and Machine Learning,
Mittweida - Germany

Abstract. Machine learning-based systems are now part of a wide array of real-world applications seamlessly embedded in the social realm. In the wake of this realisation, strict legal regulations for these systems are currently being developed, addressing some of the risks they may pose. This is the coming of age of the interpretability and explainability problems in machine learning-based data analysis, which can no longer be seen just as an academic research problem. In this tutorial, associated to ESANN 2021 special session on “Interpretable Models in Machine Learning and Explainable Artificial Intelligence”, we discuss explainable and interpretable machine learning as *post-hoc* and *ante-hoc* strategies to address these problems and highlight several aspects related to them, including their assessment. The contributions accepted for the session are then presented in this context.

1 Introduction

The design of Machine Learning (ML) models is currently dominated by the development of deep Multi-Layer Perceptrons (MLPs) and variants thereof, which consist of increasingly complex structures and modules [8, 18]. These approaches may include specific components like convolutional layers for adaptation to specific tasks like image processing and classification [32, 31]. The mathematical verification of deep networks justifies the theoretical correctness [11, 8, 18, 22].

The training of those complex models requires careful adaptation, frequently accompanied by strategies to reduce numerical instabilities, to ensure robustness and to avoid overfitting [25, 23, 47], including autoencoder learning for the pre-training of layers, dropout learning approaches, as well as regularization techniques and resilient network architectures [24, 48, 19, 32].

Nevertheless, the more complex the architectures, the more difficult the interpretation or explanation of how and why a particular network prediction is obtained, or the elucidation of which components of the complex system contributed essentially to the obtained decision. For the development of successful analyses in many application areas, this information is not demanded; yet, for

*A.V. is supported by Spanish grant PID2019-104551RB-I00.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.

many others like medical or engineering applications, especially in safety-critical contexts such as diagnostic decision support and autonomous driving, it becomes crucial to understand how the model generated the prediction. In other words, the model acting as a black-box system is not sufficient any longer [51, 53]. Importantly, for real-world applications, model interpretability or explainability or both may be even a legal requirement. In the European context, this is now enforced by the General Data Protection Regulation (GDPR) since 2018, which, as explained by Bacciu et al. [6], mandates a "right to explanation" of decisions made on citizens by "automated or artificially intelligent algorithmic systems". This is compounded by the current development of a legal framework on Artificial Intelligence (AI) by the European Commission of obvious impact on ML [16]. While mirroring many of the GDPR elements, it also discriminates AI applications according to a risk assessment, from "minimal risk" to "unacceptable risk", with high risk AI applications being subject to strict obligations before being marketed and even "limited risk" ones being tied to specific transparency obligations. Note, though, that this legal proposal often refers to *transparency* and *trustworthiness* of AI systems, instead of explainability and interpretability, but with different connotations. As argued by Fink [17], and in relation to AI explainability, Article 13 in the proposal specifies that high-risk AI systems are to be developed "to be sufficiently transparent to ensure the user's ability to interpret and use the system's output", but without including any obligations of "AI users to explain or justify the decisions they reach towards those affected by them".

This situation has made the development of tools and strategies to explain those complex models an urgent necessity [45]. As pointed out in [43], these *post hoc* strategies might be problematic, because explanations frequently are not reliable an can even be misleading. An alternative to that are interpretable models, which provide *ante hoc* inherently the possibility for model explanations.

In this tutorial, corresponding to the ESANN 2021 special session on "Interpretable Models in Machine Learning and Explainable Artificial Intelligence", we discuss both strategies in more detail and highlight several aspects related to them. We provide examples for the corresponding ML models and outline directions of ongoing and future research in this area. We also summarily describe the contributions selected for the special session.

2 Explainable, Interpretable, and Robust Models in Machine Learning

As already mentioned in the introduction, the majority of currently applied ML models are based on deep MLPs, which often achieve impressive results in regression and classification problems in very different application areas. Unfortunately, most of these complex networks work as black-box algorithms such that the user is only provided with the prediction or decision of the model, but with none or very limited information on how these results were obtained. However, the benefit of ML models will be much higher for the data analysts

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

and the experts in the application domain if they are provided with additional information about the prediction process. Such information, will increase the trustworthiness of the model, allowing the user to draw further conclusions, and extend, in this way, the knowledge base for the problem. In particular, several desiderata for robust interpretability and explainability of ML models can be identified as minimum requirements [3, 4]:

- *Explicitness and intelligibility*: Are the explanations immediately understandable?
- *Faithfulness*: Are relevance scores indicative of *true* importance?
- *Stability*: How consistent are the explanations for similar or neighbouring examples?
- *Sparsity*: Do the explanatory variables comprise, in some sense, a minimal set?
- *Transparency*: Is the model not too complex and can be decomposed in simple sub-modules which are interpretable and can be easily explained (local and global interpretability)?
- *Comprehensibility*: Is the learning approach able to represent its learned knowledge in a human understandable fashion?
- *Model inspection*: Is it possible to obtain model representations and descriptions of specific model properties?

Two main strategies in this context can be observed: *explainable artificial intelligence* (abbreviated by XAI) and *interpretable models*, which we characterize by the following definitions:

- **Explainable models**: The decision or prediction process of the model can be comprehended *post-hoc* by experts in the field using additional tools and elaborate considerations.
- **Interpretable models**: The decision or prediction process of the model can be easily comprehended by experts in the field according to the *ante-hoc* model design and their domain knowledge.

Both strategies have to provide a qualitative understanding of the process that links the input variables (features) with the outcome or response, to make the model plausible and the prediction trustable [42].

2.1 Post-hoc Approaches: Explain Machine Learning Models

Post-hoc approaches comprise those for black-box models for which explanations are sought locally to back-up individual predictions. The corresponding tools generally fall into the following five categories, starting with variants of sensitivity analysis, but extending to more complex methods:

- *Feature attribution* methods relate the model output to a small number of numeric or semantic input features. The used algorithms for this task are usually interpretable by design. However, it is difficult to derive them from data computationally efficient. Some advances have been made in generating nomograms for flexible models applied to tabular data [50].

- *Saliency maps* identify sparse components of the original signal that have most influence on the model predictions, for example, Local Interpretable Model-agnostic Explanations [42] or Class Activation Mappings [55].

- *Activation maximization*, for example, based on Generative Adversarial Networks [56], use deep generative networks and tailored optimization methods to generate class-relevant inputs for convolutional neural networks [13]. A human user can then understand the internal representations assimilated by the network and the typical representations of the classes.

- *Rule extraction based on decision trees* allows for the extraction of decision rules from deep neural networks to transfer knowledge from a reference model into an explainable equivalent [41].

- *Metric learning* consists on deriving a metric from a classifier and using it to map out the data structure [44]. Then, similarity networks are generated from which a classification of an input can be obtained by consulting its neighbours. Additionally, Siamese networks have become very popular of late in the context of self-supervised learning [37].

2.2 Ante-hoc Approaches: Interpretable Models

It has been said that the best explanation of a simple model is the model itself (i. e., it perfectly represents itself and is easy to understand). More formally, the propagation of information in a form that can be interpretable by the end-user with reasonable domain knowledge is clear from input through to prediction. Thus, interpretable models have to be transparent on all levels. The models surveyed in the following list belong to this category of transparent models:

- *Linear regression*: The linear dependencies between data and prediction make these models inherently transparent. To some extent, this concept can be adapted for logistic regression models as well.

- *Decision trees*: This rule based system generates logical implications for model prediction, i. e., it can be taken as a rule-based model. The hierarchical structure allows a transparent decision making.

- *Bayesian models*: Bayesian models form a probabilistic directed acyclic graphical model reflecting the dependencies between the input and the outcome to predicted. Recent developments include cognitive aspects of learning and knowledge representation separating detectable features and the respective reasoning for inference [46].

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.

- *Prototype methods*: These methods are based on learning of or the extraction of prototypical representations of the dataset based on a dissimilarity measures [39] and a prototype assignment rule (e. g., the nearest prototype principle, k-nearest neighbors rule). By the prototypical representations and the dissimilarity measure, this paradigm ensures interpretability in a natural manner. For classification learning, the family of Learning Vector Quantizers (LVQ) is well-known to provide possibilities for non-standard metric usage and metric adaptation [9]. The latter allows a direct evaluation of feature dependencies according to the model-inherent classification correlation analysis [52]. Unsupervised models for representation learning are the Neural Gas, Fuzzy c-Means, and Self-Organizing Maps [36, 40, 30].
- *Generalized additive models*: Recently, there has been interest in representing neural networks in the form of Generalized Additive Models, that is to say, in the form of a linear combination of interpretable non-linear functions of the inputs [2]. This approach has also been pursued with a constructive approach to infer from a trained MLP a model with univariate and bivariate effects, in the form of Partial Response Networks [33, 34]
- *Evolutionary fuzzy modelling*: Fuzzy logic systems are capable of making accurate predictions, while providing a reasonable level of interpretability [41]. This approach has been successfully applied in practical contexts including biomarker discovery and cancer diagnosis, leading to a commercial solution for the discovery of interpretable diagnostic signatures [1].
- *Self-supervised learning*: Recently, self-supervised models have become quite popular. Based on a contrastive loss, these models try to either embed the data into meaningful manifolds of low dimensionality, or to extract interpretable features, which can serve as input for other models [27, 28, 37].

2.3 How to Quantify Interpretability and Explainability

There is not yet complete consensus on how to evaluate the quality of an explainable or an interpretable method. Evaluation methods for interpretable ML include "real humans on real tasks", proposed by Doshi-Velez and Kim [14] and "AI rationalization" introduced by Ehsan et al. [15]. The quality of a given explanation needs to be evaluated in the context of its task, measuring how much the explanations facilitate and improve decision making.

A possible approach is to take an application-grounded evaluation on the respective task [10]. This requires conducting user experiments with the real applications by having the explanations tested and evaluated by the user (who is also a domain expert). A good baseline for such evaluation is how good a human would be at explaining the decisions.

Good practice in evaluating interpretability includes the following [10]:

- *Accuracy*: The actual connection between the given explanation by the explanation method and the prediction from the ML model.

- *Understandability*: This is related to the easiness with which an explanation is comprehended by the observer.
- *Efficiency*: This reflects the time necessary for a user to grasp the explanation (an explanation should be understandable in a finite and preferably short amount of time), as well as the usability of the tool presenting the explanations.

The above criteria need to take into account the *three "Cs"* of interpretability:

- *Completeness*: Verifying the validity of the explanation (i. e., the coverage of the explanation in terms of the number of instances which are comprised by the explanation).
- *Correctness*: Each explanation should generate trust. This property is related to the label coherence of the inputs covered by the explanation (i. e., inputs covered by a correct explanation should have the same label).
- *Compactness*: Each explanation should be succinct, which can be verified by the number of conditions in the decision rule and the feature dimensionality of a neighbor-based explanation.

Accordingly, an approach for model comparison and evaluation was proposed by Backhaus and Seiffert [7] based on radar plots with the three axes performance (accuracy), slimness (model complexity in terms of operations needed) and interpretability (feature weighting, class typical representations, direct decision boundaries). Another aspect suggested for the evaluation of interpretability is fidelity, understood as to which extent the model is able to imitate a black-box predictor performance compared to the black-box model itself [21].

3 Contributions from ESANN 2021

Contributions of the special session on "Interpretable Models in Machine Learning and Explainable Artificial Intelligence" cover a broad range of the previously mentioned aspects: interpretability of prototype-based methods for classification and efficient data representation [49, 20, 29], interpretability of Support Vector Machines (SVMs) [54], interpretability of random forests [38], explainability of black-box models [12, 26, 35], and informativeness of linguistic properties in word representations [5].

With respect to prototype-based models, the approach described by Kaden et al. [29] realizes information bottleneck learning by combining counterpropagation and LVQ, whereas Graeber et al. [20] uses context information and prototype adaption while inference for better LVQ performance and interpretability. Taylor and Merényi [49] propose an improvement to t-SNE which allows automated specification of its perplexity parameter using topological information about a data manifold revealed through prototype-based learning.

The partial response SVM approach proposed by Walters et al. [54] improves the explanation of feature attributing and thus contributes to the better interpretability of classification decisions generated by SVMs. In the field of random forests, a current problem is that variable importance criteria are known to be sensitive to correlated input variables, so that, for instance, the importance ranking is unreliable. Chavent et al. [38] studied this problem and present a method to estimate variable importance in the presence of correlated input variables.

In the explainability context of black-box models, Raulf et al. [12] propose a smoothed version of layer-wise relevance propagation by computing the relevance scores as averages over noisy inputs. The conducted experiments show that the smoothed version leads to improved explanations. Moreover, Izzo et al. [26] studies the determination of the Shapley baseline value because an inappropriate choice of baseline could negatively impact the explanatory power of the method and possibly lead to incorrect interpretations. To avoid such defects, they present a method for choosing a baseline according to a neutrality value that is in accordance with how the model is used while decision making. Furthermore, Madhikermi et al. [35] propose an adaptive weighted sampling method that improves the representativeness of the generated samples in the presence of strong non-linearities or exceptional input feature value combinations. To verify this sampling strategy, they integrated it into the calculation approaches of contextual importance and utility of features, which are sampling sensitive explanation methods.

Finally, Babazhanova et al. [5] studies the informativeness of linguistic properties such as part-of-speech and named entities encoded in word representations and show that the part-of-speech information is more important for word embeddings than the named entity property.

4 Conclusions

AI and, central to it, ML, are becoming increasingly bound by law, regulatory frameworks, and ethical guidelines, all of which, in one way or another, place their focus on issues of algorithmic trustworthiness, transparency, interpretability, and explainability. Much of this responsibility is placed on the shoulders of the data controllers and data analysts, which implies that the methods required to comply with these obligations must enter a phase of maturity.

In this brief paper, we have described explainability and interpretability as *post hoc* and *ante hoc* strategies. The former strategies can be seen as depending on developments beyond the modeling itself (often domain-specific), whereas the latter strategies focus on interpretable models that inherently offer the possibility for model explanations.

References

[1] Precision quartzbio^{SM}, 2021.

[2] R. Agarwal, N. Frosst, X. Zhang, R. Caruana, and G. Hinton. Neural additive models: Interpretable machine learning with neural nets. *arXiv*, 2004.13912, 2020.

[3] D. Alvarez-Melis and T. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS 2018)*, pages 7786–7795, 2018.

[4] A. Arrieta, N. Díaz-Rodríguez, J. D. Serac, A. Bennetot, S.Tabikg, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[5] M. Babazhanova, M. Tezekbayev, and Z. Assylbekov. Geometric probing of word vectors. In M. Verleysen, editor, *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2020), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2021. i6doc.com.

[6] D. Bacciu, B. Biggio, P. Lisboa, J. Martín, L. Oneto, and A. Vellido. Societal issues in machine learning: When learning from data is not enough. In *Proceedings of the 27^{th} European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2019)*, pages 455–464, 2019.

[7] A. Backhaus and U. Seiffert. Classification in high-dimensional spectral data: Accuracy vs. interpretability vs. model size. *Neurocomputing*, 131:15–22, 2014.

[8] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[9] M. Biehl, B. Hammer, and T. Villmann. Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2):92–111, 2016.

[10] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[11] G. Cybenko. Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[12] S. Däubner, A. Raulf, B. Hack, A. Mosig, and A. Fischer. SmoothLRP: Smoothing LRP by averaging over stochastic input variations. In M. Verleysen, editor, *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2020), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2021. i6doc.com.

[13] J. Despraz, S. Gomez, H. F. Satizábal, and C.-A. Peña-Reyes. Towards a better understanding of deep neural networks representations using deep generative networks. In *Proceedings of the 9th International Joint Conference on Computational Intelligence (IJCCI 2017)*, pages 215–222. SCITEPRESS – Science and Technology Publications, 2017.

[14] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv*, 1702.08608v2, 2017.

[15] U. Ehsan, B. Harrison, L. Chan, and M. Riedl. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES'18)*, pages 81–87, 2018.

[16] European Commission. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206`, 2021.

[17] M. Fink. The EU Artificial Intelligence Act and access to justice. *EU Law Live*, 2021.

[18] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, Cambridge, MA, 2016.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680, San Diego, 2014. Curran Associates, Inc.

[20] T. Graeber, S. Vetter, S. Saralajew, M. Unterreiner, and D.Schramm. AGLVQ - Making generalized vector quantization algorithms aware of context. In M. Verleysen, editor, *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2020), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2021. i6doc.com.

[21] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, and F. Giannotti. A survey of methods

for explaining black box models. *ACM Computing Surveys*, 51(5/93):1–42, 2019.

[22] B. Hanin. Universal function approximation by deep neural networks with bounded width and ReLU activations. *Mathematics*, 7(992):1–9, 2019.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*, pages 770–778, 2016.

[24] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(7):5, 2006.

[25] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[26] C. Izzo, A. Lipani, R. Okhrati, and F. Medda. A baseline for shapley values in MLPs: From missingness to neutrality. In M. Verleysen, editor, *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2020), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2021. i6doc.com.

[27] T. Jakub, A. Gupta, H. Bilen, and A. Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8787–8797, 2020.

[28] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–22, 2021.

[29] M. Kaden, R. Schubert, M. Mohannazadeh-Bakhtiari, L. Schwarz, and T. Villmann. The LVQ-based counter propagation network an interpretable information bottleneck approach. In M. Verleysen, editor, *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2020), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2021. i6doc.com.

[30] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[31] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 1097–1105. Curran Associates, Inc., San Diego, 2012.

[32] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 52:1097–1105, May 2015.

[33] P. Lisboa, S. Ortega-Martorell, S. Cashman, and I. Olier. The partial response network: a neural network nomogram. *arXiv*, 1908.05978, 2019.

[34] P. Lisboa, S. Ortega-Martorell, and I. Olier. Explaining the neural network: A case study to model the incidence of cervical cancer, proceedings part iii. In M.-J. Lesot, S. Vieira, M. Reformat, J. P. Carvalho, A. Wilbik, and B. B.-M. R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 99–113, 2020.

[35] M. Madhikermi, A. Malhi, and K. Främling. Context-specific sampling method for contextual explanations. In M. Verleysen, editor, *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2020), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2021. i6doc.com.

[36] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569, 1993.

[37] I. Misra and L. van Maaten. Self-supervised learning of pretext-invariant representations. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6706–6716. IEEE Press, 2020.

[38] A. Mourer, M. Chavent, M. Olteanu, and J. Lacaille. Handling correlations in random forests: which impacts on variable importance and model interpretability? In M. Verleysen, editor, *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2020), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2021. i6doc.com.

[39] D. Nebel, M. Kaden, A. Villmann, and T. Villmann. Types of (dis−)similarities and adaptive mixtures thereof for improved classification learning. *Neurocomputing*, 268:42–

54, 2017.

[40] N. R. Pal, J. C. Bezdek, and R. J. Hathaway. Sequential competitive learning and the fuzzy c-means clustering algorithms. *Neural Networks*, 9(5):787–796, 1996.

[41] C.-A. Peña-Reyes and M. Sipper. *Accuracy Improvements in Linguistic Fuzzy Modeling*, volume 129 of *Studies in Fuzziness and Soft Computing*, chapter Fuzzy CoCo: Balancing Accuracy and Interpretability of Fuzzy Models by Means of Coevolution, pages 119–146. Springer, 2003.

[42] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, pages 1135–1144, 2016.

[43] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[44] H. Ruiz, T. A. Etchells, I. H. Jarman, J. Martin, and P. J. G. Lisboa. A principled approach to network-based classification and data representation. *Neurocomputing*, 12(7):79–91, 2013.

[45] W. Samek, G. Monatvon, A. Vedaldi, L. Hansen, and K.-R. Müller, editors. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, number 11700 in LNAI. Springer, 2019.

[46] S. Saralajew, L. Holdijk, M. Rees, E. Asan, and T. Villmann. Classification-by-components: Probabilistic modeling of reasoning over a set of components. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, pages 2788–2799. MIT Press, 2019.

[47] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[48] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[49] J. Taylor and E. Merényi. A parameterless t-SNE for faithful cluster embeddings from prototype-based learning and CONN similarity. In M. Verleysen, editor, *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2020), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2021. i6doc.com.

[50] V. van Belle, B. van Calster, S. van Huffel, J. A. K. Suykens, and P. Lisboa. Explaining support vector machines: A color based nomogram. *PLOS ONE*, 11(10):e0164568, 2016.

[51] A. Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 32:18069–18083, 2020.

[52] T. Villmann, A. Bohnsack, and M. Kaden. Can learning vector quantization be an alternative to SVM and deep learning? *Journal of Artificial Intelligence and Soft Computing Research*, 7(1):65–81, 2017.

[53] T. Villmann, S. Saralajew, A. Villmann, and M. Kaden. Learning vector quantization methods for interpretable classification learning and multilayer networks. In C. Sabourin, J. Merelo, A. Barranco, K. Madani, and K. Warwick, editors, *Proceedings of the 10th International Joint Conference on Computational Intelligence (IJCCI), Sevilla*, pages 15–21, Lissabon, Portugal, 2018. SCITEPRESS - Science and Technology Publications, Lda. ISBN: 978-989-758-327-8.

[54] B. Walters, S. Ortega-Martorell, I. Olier, and P. Lisboa. The partial response SVM. In M. Verleysen, editor, *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2020), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2021. i6doc.com.

[55] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[56] Z. Zhou, S. Rong, Y. Song, K. Ren, J. Wang, W. Zhang, and Y. Yong. Activation maximization generative adversarial networks. In *Proceeding of the 6th International Conference on Learning Representations (ICLR 2018)*, pages 1–24, 2018.