Explainable Artificial Intelligence: An Overview on Hybrid Models

Gabriel Quesada[1],*, María José del Jesus[1] and Pedro González[1]

[1]Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Jaén, 23071, Jaén, Spain

Abstract

The increasing integration of Artificial Intelligence (AI) in various critical areas highlights the need to both achieve accuracy in predictions and understand the logic behind them for proper decision making. Explainable Artificial Intelligence (XAI) addresses this challenge, balancing the complexity of models with the necessary transparency and interpretability. Hybrid models, by integrating the accuracy of black-box models with the transparency of interpretable ones, represent a promising avenue in the move towards more understandable, accurate and reliable systems in AI, encouraging their safe, ethical and responsible adoption in diverse real-world applications. This paper provides an exploration of hybrid models in XAI, elaborating on key concepts and offering a classification based on interpretability. In addition to describing the construction of these models, it reviews advances in the literature and identifies future directions.

Keywords

Hybrid Models, Explainable Artificial Intelligence, XAI, Interpretability, Black-Box Models

1. Introduction

The increasing adoption of Artificial Intelligence (AI) systems in a wide range of critical applications has highlighted the importance of not only achieving high levels of prediction accuracy, but also of understanding and explaining the reasoning behind these decisions. In this context, Explainable Artificial Intelligence (XAI) has emerged as a crucial area of research that seeks to balance the complexity of AI models with the need for transparency and interpretability.

On the one hand, interpretable models have gained relevance, as they allow researchers and practitioners to extract knowledge directly from model outputs. However, these models often sacrifice complexity and performance in favor of transparency [1]. On the other hand, black-box models, such as deep neural networks, have demonstrated outstanding performance on a variety of complex tasks. However, their lack of interpretability can suppose significant challenges in environments where explainability is required, such as in medical, financial or legal applications [2]. In response to this dichotomy, a promising approach has emerged: *hybrid models*, which combine the best of both worlds by integrating elements of interpretable models and black-box ones. These models offer a balance between performance and understanding, allowing both

MAI-XAI'24: Workshop "Multimodal, Affective and Interactive eXplainable Artificial Intelligence", October 20, 2024, Santiago de Compostela, Spain

*Corresponding author.

gquesada@ujaen.es (G. Quesada); mjjesus@ujaen.es (M. J. d. Jesus); pglez@ujaen.es (P. González)

0009-0007-3653-3334 (G. Quesada); 0000-0002-7891-3059 (M. J. d. Jesus); 0000-0002-6733-3868 (P. González)

accuracy in prediction and the ability to interpret [3] and justify those predictions, which is crucial in XAI [4].

In this article, an analysis of the hybrid models present in the literature is performed with the aim of justifying their relevance and necessity in XAI. In addition, the focus will be on identifying the current challenges faced by these hybrid models and outlining the lines of future work needed to address them. Through this detailed analysis, a comprehensive view of the essential elements for the development and improvement of new hybrid algorithms is aimed to be provided, thus contributing to the continued advancement of research in this area of AI.

The remainder of the paper is organized as follows: section 2 presents the main concepts related to XAI, as well as a classification of AI models according to their interpretability. In section 3, the construction of hybrid models and their state of the art on the models of this type existing in the literature, including Interpretable Modeling Techniques Approaches and Neural-Symbolic ones, are presented. Finally, in section 4, the open lines of research in this area and the conclusions drawn from this work are indicated.

2. Background

In this section, three main aspects of XAI will be explored: explainability, interpretability and transparency. First, these concepts will be defined, highlighting their importance and their relationship to the development of AI systems that can be understood and trusted by human users. Subsequently, a XAI classification based on model interpretability will be introduced, allowing us to group and better understand the approaches and techniques used in the area.

2.1. Explainability, interpretability and transparency

XAI emerges as crucial in the area of AI, where understanding and trusting AI systems is imperative. XAI focuses on developing methods and techniques that enable users to understand and trust the decisions made by AI models. In this context, explainability, interpretability and transparency play a key role, as they are intrinsically linked in the development of AI systems that are ethical, trustworthy and socially responsible, and are fundamental to addressing concerns about opacity and bias in algorithmic decision making [5].

In XAI, explainability, interpretability and transparency terms refer to the ability to understand, explain and provide visibility into the operation and decisions of an AI model in a way that is clear and understandable to humans [5]. Although these three concepts are closely related and are often used interchangeably, they have different meanings:

- *Explainability*: It refers to "producing details or reasons to make its functioning clear or easy to understand" [6] for a given audience. This can be crucial for users to accept and trust the recommendations or decisions of an AI system.
- *Interpretability*: It denotes "the ability to understand the internal mechanics of a machine learning model. It denotes the extent to which a human can comprehend the cause of a decision"[7]. An interpretable model allows users to understand how decisions were made and what features or factors influenced those decisions.

- *Transparency*: This concerns the degree to which the decision-making process of an AI system is clear and understandable to users, allowing critical evaluation by users and stakeholders.

In summary, while explainability focuses on the system's ability to provide clear explanations, interpretability relates to the user's ability to understand the model itself, and transparency relates to the clarity and understanding of the decision-making process of the system as a whole [8]. All three aspects are important for the development and adoption of trustworthy AI systems.

2.2. Classification of AI models according to interpretability

The classification previously established in [4] [9] provides a framework for understanding how different approaches in XAI address the challenge of model interpretability, from fully interpretable models to inherently opaque ones:

1. *White-box models* [10]: They are AI models whose internal structure and operation are completely transparent and understandable to users. These models allow an easy interpretation of how decisions are made and how predictions are generated. Depending on whether or not they present some small difficulty to be interpreted, they can be classified into the following two types [11]: *fully interpretable* ones, which provide a complete understanding of how decisions are made, and are usually simple and easily understandable, such as linear regressions, decision trees and association rules; and *partially interpretable* ones, which may require effort to fully understand how they work, such as logistic regressions.
2. *Black-box models* [8] [12]: These are opaque and lack inherent interpretability due to their complexity and lack of transparency. Examples include deep neural networks, support vector machines (SVMs), random forests or Gradient Boosting Machines. While these models can achieve high predictive performance, their lack of interpretability can pose significant challenges in terms of explainability and confidence for users.
3. *Gray-box models* [4]: This type of model discloses its internal structure and workings. Unlike black-box models, which are completely opaque, and white-box models, which are fully transparent, gray-box models reveal some aspects of their internal logic and processes. This approach provides a balance between reliability and comprehensibility of the models [9]. Within this type of models we distinguish the following two subtypes:
    - *Hybrid models* [4] [3]: These models combine elements of both transparent (white-box) and opaque (black-box) models. These models aim to leverage the high performance of black-box models while incorporating the interpretability of white-box models, offering a balance between accuracy and explainability. They are designed to provide insights into the decision-making process while maintaining robust predictive capabilities.
    - *Explanatory models* [4]: This type of model focus on generating comprehensible explanations for black-box models, without significantly modifyigng their internal structure. They use techniques such as weighted attention or feature relevance to highlight the most influential parts of the data in model decision making.

In the next section, we will focus on addressing the challenges of explainability in AI through hybrid models due to their ability to combine the strengths of white-box and black-box models, their flexibility and adaptability to different needs, contexts and levels of explainability, and their robustness and generability by combining different approaches and techniques.

3. Overview of hybrid models

In this section, an exploration of a variety of hybrid approaches that combine elements of interpretable models and black-box ones to address XAI challenges will be undertaken. The discussion commences with an examination of how these hybrid models are constructed, highlighting the different approaches and methodologies employed to integrate explainability into AI systems. Subsequently, attention will be directed towards interpretable modeling techniques and approaches, delving into how these methods can complement and enhance the understanding of more complex AI models. Finally, neural-symbolic approaches will be explored, wherein the deep learning capabilities of neural networks are combined with symbolic logic and human reasoning, offering a unique perspective on interpretability in AI. Through these subsections, a comprehensive and detailed overview of hybrid models in the creation of XAI systems is aimed to be provided.

3.1. How to build hybrid models

Beginning with a discussion on hybrid models requires explaining how they are built, as it offers insight into the steps, challenges, and decisions involved in integrating interpretable and deep learning components.

A framework is proposed [13] for building a hybrid predictive model from an interpretable model and any pre-trained black-box model, in such a way as to combine their strengths: the transparency of the first with the precision of the second. To accomplish this, the interpretable model would replace the black-box model on a subset of the data where it demonstrates accuracy: an input is first sent to the interpretable model to see if a prediction can be generated directly and, if not, the black-box model is activated. This partitioning of the data set can be done by sets of association rules or linear models using a threshold and evaluation of the trade-off between transparency and accuracy through Pareto frontiers.

Within the realm of constructing hybrid models, an exploration reveals three distinct paradigms [3], each characterized by the timing of training for the model components. These paradigms offer varying approaches to the integration of interpretable and black-box model elements, showcasing different strategies for achieving the desired balance between transparency and performance:

1. *Post-Black-Box Paradigm*: It consists of first training a highly accurate black box model and then wrapping it with an interpretable model, thus increasing its transparency and correcting the errors made by the black box as its predictions are known in advance.
2. *Pre-Black-Box Paradigm*: It comprises learning the interpretable part of the model before training a black box model with the remaining examples. First, the easiest examples are

identified and a simple model is trained with them. Then, the black box part will be used to classify the examples not sent to the simple part in order to increase performance.

3. *End-to-End Approach*: It involves simultaneously training both components of a hybrid model. Although this approach would theoretically produce the best results by guaranteeing global optimality, it is also very challenging, as it requires coding both the interpretable and the black-box models within the same framework. To our knowledge, this proposal has not yet been addressed in the literature.

In practice, the construction of hybrid models predominantly follows the *Post-Black-Box Paradigm*. This paradigm aligns with the methodologies employed in both the *Interpretable Modeling Techniques Approaches* and *Neural-Symbolic Approaches* subsections that follow. The first one focuses on interpretable techniques, such as linear regression or decision trees, while the second one concentrates on neural-symbolic approaches that combine deep learning capabilities with symbolic representation.

3.2. Interpretable Modeling Techniques Approaches

One of the prevailing methodologies in this domain involves integrating Deep Neural Networks (DNN) with inherently interpretable modeling techniques [4]. This approach seeks to combine the powerful predictive capabilities of DNNs with the inherent transparency and comprehensibility offered by specific modeling methods, such as decision trees and linear regression. By leveraging interpretable methods alongside DNNs, researchers aim to develop hybrid models that not only deliver high predictive accuracy but also provide insights into the underlying decision-making process, thus enhancing trust and understanding in AI systems.

One proposal that is quite generalizable because it does not require any particular interpretable model, is *Greybox XAI* [14], which consists of two separately trained models: a DNN whose purpose is to detect the different parts of the object that constitute an image, and a transparent model that encodes the presence and absence of parts of the object. One of the main interpretable algorithms that can be combined with a DNN is k-Nearest Neighbors (kNN), giving rise to the *Deep k-Nearest Neighbors* (DkNN) algorithm [15], by employing its inference on the latent representation of the training data set acquired through layers of a DNN. DKNNs have proven to be efficient and robust, and offer example-based and neighborhood-based explanations. Another model that applies neighborhood information and ensemble methods is *Deep Weighted Averaging Classifiers* (DWAC) [16], which combines the output of multiple classifiers using a weighted average, where weights are learned adaptively. This model attempts to reduce the disparity in predictions between different groups taking into account the weights of instances based on their features. While DWAC uses a weighted average approach to calculate the weight of each classifier, another hybrid model, *Bayesian deep learning* (BDL) [17], estimates the uncertainty of black-box models by putting distributions on model weights or by looking for a direct mapping of probabilistic outputs. Another example of a model using a linear classifier would be the *Self-Explaining Neural Networks* (SENN) [18], generalizing it by using neural networks to learn their characteristics, the coefficients associated with them and how the networks are aggregated to make a prediction. Also focused on classification tasks and generating predictions through deep neural networks that can be interpretable are *Contextual Explanation Networks* (CEN)

[19], although these focus on contextualizing classification decisions through a contextual classification network and an explanation generator.

We also find within hybrid models those that combine DNN with memory structures, known as *Memory networks* [20], which are a type of neural network architecture that includes a mechanism for storing information mimicking the short-term memory of the human brain. A variant of this type of networks is *BagNets* [21], which combines a convolutional network with a memory structure to store contextual information during input reading. Another system that uses a convolutional neural network as one of its components is the *Hybrid Convolutional Fuzzy Classifier* [22], which uses it as a feature extractor, and then fuzzy clustering is applied for classifying the extracted features. Another approach using a fuzzy logic classifier is the *Self-configuring evolutionary algorithms* [23], in this case combined with an artificial neural network (ANN), which builds a rule base interpreted from the inputs and outputs of the ANN. Continuing with the algorithms that use the clustering technique, we have *Cluster-TREPAN* [24], which proposes the combination of two methods to explain the predictions given by a neural network: the TREPAN algorithm, which produces a decision tree that approximates the function represented by the network; and a hidden-layer clustering for neural networks, which analyzes the causal importance of features at the cluster level.

Additionally, hybrid models can be built from an additive perspective. On the one hand, there is the *Generalized Additive Neural Network Model* (GANNM) [25], which combines the high-precision predictive performance of neural networks with interpretability and flexible function forms of Generalized Additive Models (GAMs) to model complex relationships in structured data. On the other hand, we find *Adaptive explanatory neural networks* (AxNNs) [26], which consist of a two-stage approach, one using a network of GAMs to capture the main effects and the other through an explanatory neural network. Continuing with the adaptive models, we find the *Locally Adaptive Interpretable Regression* [27], which consists of two main components: a linear regression model, which uses Ordinary Least Squares to obtain the regression coefficients and their standard errors, and a learning model parameterized by deep neural networks that predicts the percentiles of a Gaussian distribution for the regression coefficients. Another adaptive model is *AdaAX* [28], which integrates Recurrent Neural Networks (RNN) with Adaptive State Learning Automata, modeling and explaining the decisions made by the RNN.

Another system that uses RNN is the *Explanations via Model Extraction* [29], which consists of a model extraction approach capable of approximating RNNs with interpretable models represented by human-understandable concepts and their interactions, such as K-means clustering. In turn, within the RNNs we can find a more specific type of networks, the LSTM (Long Short-Term Memory), which thanks to their capacity to retain and remember long-term information makes them ideal for combining with ARIMA, the traditional statistical models for time series forecasting, creating a hybrid, *ARIMA-LSTM* [30], which is capable of detecting both linear and nonlinear trends in the model.

3.3. Neural-Symbolic Approaches

In addition, hybrid models have also explored the integration of symbolic learning techniques, such as decision rules or fuzzy logic, with data-driven machine learning models. This combination allows for capturing expert knowledge or heuristic rules in the model, which facilitates its interpretation and understanding by end users. These algorithms are the so-called *Neuro-symbolic models* [31], which apply connectionist mechanisms to the principles of computation, characterization and analysis of symbolic computation to improve the interpretability of the neural network. Among this type of algorithms are *Conceptors* [32], which restrict the representations learned by a deep neural network to a specific concept, being applied as an additional layer at the end of the network that makes them more stable, interpretable and robust to perturbations, allowing them to generalize to new data and increase their explainability. With a similar objective, *Logic-based concept induction* can be applied [33], which combines complex machine learning models, usually neural networks, with the extraction of interpretable logical rules from the decisions made by the model; or the *Logic Explained Networks* [34], which, once a neural network has been trained, generates a set of interpretable explanations in the form of logical rules or propositions that describe the patterns identified by the model, integrating them into the model so that they can be associated with the predictions made by the neural network.

Following this same NeSy paradigm, some models can be found that go beyond explainability, identifying differences with expected explanations and correcting them. This is the case of *X-NeSyL* [35], which aligns the representation of deep learning, symbolic logic and domain expert knowledge through an explainability feedback mechanism. In this way, explainable, theory-driven data science is achieved. Continuing with the idea that having prior knowledge about the environment can improve explainability, we find the *PLENARY* [36] algorithm, which generates linguistic summaries describing the features and patterns that the black box model uses to make decisions. These summaries are expressed in a natural language understandable to humans and provide an intuitive explanation of how the model makes its predictions.

4. Conclusions

XAI is a crucial area of research that is rapidly evolving to address the challenges of trust and ethics in AI. Interpretable hybrid models are a promising research area with implications in a wide range of fields by combining the feature representation capability of deep neural networks with the interpretability of transparent systems. The research in this area can lead to more interpretable machine learning systems with the goal of making AI more transparent, reliable, and understandable to all stakeholders. In this paper, a closer look at the process of building hybrid models is addressed, providing detailed information on their architecture and methodologies. The different paradigms under which hybrid models can be developed depending on the order in which their constituent components are trained are also shown. In addition, the hybrid explainable models documented in the literature are reviewed.

The research carried out reveals interesting lessons learned. When building hybrid models of this kind it is necessary to achieve an optimal balance in maximizing both accuracy and interpretability, also generating human-understandable explanations for the predictions obtained by the model. As a consequence, confidence and usefulness in real-world applications are improved, especially those for which understanding of decision-making is critical. Furthermore, the importance of integrating different techniques in the construction of hybrid models is emphasized as it allows addressing the limitations and challenges that single approaches often present. To this end, various methods and strategies can be explored, which combine predictive

—including rule-based— probabilistic and machine learning models in a common framework.

In future research on hybrid models, priority could be given to the development of methods and tools that provide comprehensible explanations of how model predictions are obtained, allowing their application in a wide variety of areas and real-world situations, especially in those where interpretability is crucial for decision making, such as healthcare or security. On the other hand, it could be explored how to obtain more complete, versatile and robust systems by integrating hybrid models with other machine learning techniques, using ensemble or transfer learning methods, in such a way that they can satisfy a wider variety of needs, domains and data types. Finally, it is necessary to develop specific metrics to evaluate and validate the confidence, robustness and uncertainty of the predictions made by hybrid models, as well as the impact of the explanations generated and their acceptance by users.

As future work, we plan to extend and improve the overview proposed in this paper. This will include, from a methodological point of view, a systematic review of the literature on hybrid models, as well as a more in-depth approach to real-world applications of this type of models.

References

[1] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. doi:10.1038/s42256-019-0048-x.

[2] Y. B. Yann LeCun, G. Hinton, Deep learning, Nature 521 (2015) 436–444. doi:10.1038/nature14539.

[3] J. Ferry, G. Laberge, U. Aïvodji, Learning hybrid interpretable models: Theory, taxonomy, and methods, arXiv preprint (2023). doi:10.48550/arXiv.2303.04437.

[4] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. D. Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, Information Fusion 99 (2023) 101805. doi:10.1016/j.inffus.2023.101805.

[5] N. Díaz-Rodríguez, J. D. Ser, M. Coeckelbergh, M. L. de Prado, E. Herrera-Viedma, F. Herrera, Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation, Information Fusion 99 (2023) 101896. doi:10.1016/j.inffus.2023.101896.

[6] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115. doi:10.1016/j.inffus.2019.12.012.

[7] A. Hanif, A. Beheshti, B. Benatallah, X. Zhang, Habiba, E. Foo, N. Shabani, M. Shahabikargar, A comprehensive survey of explainable artificial intelligence (XAI) methods: Exploring transparency and interpretability, in: Web Information Systems Engineering – WISE 2023, Springer, Cham, 2023, pp. 915–925. URL: https://link.springer.com/chapter/10.1007/978-3-031-21657-4_76. doi:10.1007/978-3-031-21657-4_76.

[8] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[9] J. M. Alonso, C. Castiello, L. Magdalena, C. Mencar, Explainable Fuzzy Systems: Paving the way from Interpretable Fuzzy Systems to Explainable AI Systems, volume 970 of *Studies in Computational Intelligence*, Springer International Publishing, 2021. doi:10.1007/978-3-030-71098-9.

[10] C. Molnar, Interpretable Machine Learning. A guide for making black box models explainable, 2 ed., leanpub.com, 2022.

[11] B. Mittelstadt, Interpretability and transparency in artificial intelligence, in: The Oxford Handbook of Ethics of AI, Oxford University Press, 2022, pp. 378–410. doi:10.1093/oxfordhb/9780198857815.013.20.

[12] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (2018) 93. doi:10.1145/3236009.

[13] T. Wang, Q. Lin, Hybrid predictive models: When an interpretable model collaborates with a black-box model, Journal of Machine Learning Research 22 (2021) 1–38. URL: http://jmlr.org/papers/v22/19-325.html.

[14] A. Bennetot, G. Franchi, J. D. Ser, R. Chatila, N. Díaz-Rodríguez, Greybox XAI: A neural-symbolic learning framework to produce interpretable predictions for image classification, Knowledge-Based Systems 258 (2022) 109947. doi:10.1016/j.knosys.2022.109947.

[15] L. Le, Y. Xie, V. V. Raghavan, Deep similarity-enhanced K nearest neighbors, in: 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 2643–2650. doi:10.1109/BigData.2018.8621894.

[16] D. Card, M. Zhang, N. A. Smith, Deep weighted averaging classifiers, in: ACM FAT* Conference 2019, Association for Computering Machinery, 2019, pp. 369–378. doi:10.1145/3287560.3287595.

[17] A. G. Wilson, P. Izmailov, Bayesian deep learning and a probabilistic perspective of generalization, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 4697–4708. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/322f62469c5e3c7dc3e58f5a4d1ea399-Paper.pdf.

[18] D. Alvarez-Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks, in: 32nd Conference on Neural Information Processing Systems (NIPS 2018), Neural Information Processing Systems Foundation, 2018, pp. 7786–7795. URL: https://dl.acm.org/doi/10.5555/3327757.3327875.

[19] M. Al-Shedivat, A. Dubey, E. Xing, Contextual explanation networks, Journal of Machine Learning Research 21 (2020) 7950–7993. URL: https://dl.acm.org/doi/abs/10.5555/3455716.3455910.

[20] S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, End-to-end memory networks, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, MIT Press, 2015, pp. 2440––2448. URL: https://dl.acm.org/doi/10.5555/2969442.2969512.

[21] W. Brendel, M. Bethge, Approximating cnns with bag-of-local-features models works surprisingly well on imagenet, in: Seventh International Conference on Learning Representations (ICLR 2019), 2019. URL: https://openreview.net/forum?id=SkfMWhAqYQ.

[22] M. Yeganejou, S. Dick, J. Miller, Interpretable deep convolutional fuzzy classifier, IEEE Transactions on Fuzzy Systems 28 (2020) 1407–1419. doi:10.1109/TFUZZ.2019.

2946520.

[23] P. A. Sherstnev, Self-configuring evolutionary algorithms based design of hybrid interpretable machine learning models, in: International Workshop Hybrid methods of modeling and optimization in complex systems, 2022, pp. 313–321. doi:10.15405/epct.23021.38.

[24] T. De, P. Giri, A. Mevawala, R. Nemani, A. Deo, Explainable AI: A hybrid approach to generate human-interpretable explanation for deep learning prediction, Procedia Computer Science 168 (2020) 40–48. doi:10.1016/j.procs.2020.02.255.

[25] T. Wang, C. He, F. Jin, Y. J. Hu, Evaluating the effectiveness of marketing campaigns for malls using a novel interpretable machine learning model, Information Systems Research 33 (2022) 659–677. doi:10.1287/isre.2021.1078.

[26] J. Chen, J. Vaughan, V. Nair, A. Sudjianto, Adaptive explainable neural networks (AxNNs), SSRN Electronic Journal (2020). doi:10.2139/ssrn.3569318.

[27] L. Munkhdalai, T. Munkhdalai, V.-H. Pham, M. Li, K. H. Ryu, N. Theera-Umpon, Recurrent neural network-augmented locally adaptive interpretable regression for multivariate time-series forecasting, IEEE Access 10 (2022) 11871–11885. doi:10.1109/ACCESS.2022.3145951.

[28] D. Hong, A. M. Segre, T. Wang, AdaAX: Explaining recurrent neural networks by learning automata with adaptive states, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, 2022, pp. 574––584. doi:10.1145/3534678.3539356.

[29] A. Yan, T. Huang, L. Ke, X. Liu, Q. Chen, C. Dong, Explanation leaks: Explanation-guided model extraction attacks, Information Sciences 632 (2023) 269–284. doi:10.1016/j.ins.2023.03.020.

[30] S. Nie, Q. Liu, H. Ji, R. Hong, S. Nie, Integration of ARIMA and LSTM models for remaining useful life prediction of a water hydraulic high-speed on/off valve, Applied Sciences 12 (2022) 8071. doi:10.3390/app12168071.

[31] D. J. Agravante, D. Kimura, M. Tatsubori, A. Munawar, A. Gray, Learning neuro-symbolic world models with conversational proprioception, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 648–656. doi:10.18653/v1/2023.acl-short.57.

[32] H. Jaeger, Using conceptors to manage neural long-term memories for temporal patterns, Journal of Machine Learning Research 18 (2017) 1–43. URL: http://jmlr.org/papers/v18/15-393.html. doi:10.5555/3122009.3122022.

[33] M. K. Sarker, P. Hitzler, Efficient concept induction for description logics, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 3036–3043. doi:10.1609/aaai.v33i01.33013036.

[34] G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Liò, M. Maggini, S. Melacci, Logic explained networks, Artificial Intelligence 314 (2023) 103822. doi:10.1016/j.artint.2022.103822.

[35] N. Díaz-Rodríguez, A. Lamas, G. Franchi, I. Donadello, S. Tabik, D. Filliat, P. Cruz, R. Montes, F. Herrera, Explainable neural-symbolic learning (X-NeSyL) methodology to fuse deep learning representations with expert knowledge graphs: The monuMAI cultural heritage use case, Information Fusion 79 (2021) 58–83. doi:10.1016/j.inffus.2021.09.022.

[36] K. Kaczmarek-Majer, G. Casalino, G. Castellano, M. Dominiak, O. Hryniewicz, O. Kamińska, G. Vessio, N. Díaz-Rodríguez, PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries, Information Sciences 614 (2022) 374–399. doi:10.1016/j.ins.2022.10.010.