Explainable Artificial Intelligence: An Updated Perspective

Conference Paper · May 2022

DOI: 10.23919/MIPRO55190.2022.9803681

4 authors:

Agneza Krajna
University of Zagreb Faculty of Electrical Engineering and Computing
13 PUBLICATIONS **96** CITATIONS
SEE PROFILE

Mihael Kovač
University of Zagreb
13 PUBLICATIONS **45** CITATIONS
SEE PROFILE

Mario Brcic
University of Zagreb
50 PUBLICATIONS **1,296** CITATIONS
SEE PROFILE

Ana Šarčević
University of Zagreb Faculty of Electrical Engineering and Computing
8 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Explainable Artificial Intelligence: An Updated Perspective

Agneza Krajna*, Mihael Kovac*, Mario Brcic*, Ana Šarčević*
*University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
agneza.krajna@fer.hr, mihael.kovac@fer.hr, mario.brcic@fer.hr, ana.sarcevic@fer.hr

***Abstract*—Artificial intelligence has become mainstream and its applications will only proliferate. Specific measures must be done to integrate such systems into society for the general benefit. One of the tools for improving that is explainability which boosts trust and understanding of decisions between humans and machines. This research offers an update on the current state of explainable AI (XAI). Recent XAI surveys in supervised learning show convergence of main conceptual ideas. We list the applications of XAI in the real world with concrete impact. The list is short and we call to action - to validate all the hard work done in the field with applications that go beyond experiments on datasets, but drive decisions and changes. We identify new frontiers of research, explainability of reinforcement learning and graph neural networks. For the latter, we give a detailed overview of the field.**

***Index Terms*—artificial intelligence, explainability, interpretability, AI safety, graph neural networks**

I. Introduction

Artificial intelligence (AI) is one of the key drivers of industrial development today. Ethical concerns emerged only recently with the advent of model predictions that can and do cause harmful consequences. There are examples of a bias towards gender and race in AI models used in criminal justice for criminal decisions [1] and job selection [2]. In order to prevent such difficulties and finally take care of safety, the goal is to establish a regulatory or certification body that will issue permits for the commissioning of artificially intelligent systems [3] and education programs addressing the issue have been commenced [4]. Due to these problems, a relatively new branch of science, AI safety, has developed [5]. It deals with issues of security and the usefulness of AI for people and civilization in general. For people to be sure that AI systems make decisions that are not harmful according to human views, it is necessary to get an explanation for making those decisions. So, the goal of explainable artificial intelligence (XAI) is to ensure that algorithm decisions, as well as the data that influenced such a decision, can be understood by the end-users without the use of cluttered or alien expressions and terms [6], [7].

In section II we give a review of notable survey papers from 2020 onward with their contributions. Section III lists the types of use of XAI techniques and examples of concrete applications that have had a significant impact in some specific domain or on the business of a particular organization. Section IV gives an overview of the current state of XAI for Graph Neural Networks (GNNs). We conclude in section V.

II. Survey papers on XAI techniques

The manner and style of interpretation might vary depending on the problem's specific domain (e.g. medical domain, economic domain, etc.). There are a number of mechanisms by which the machine model of knowledge approaches the human model. Some examples of such mechanisms are concrete images, textual descriptions, diagrams, and understandable graphical representations of the decision-making process such as a decision tree or a specific example close to the human [8].

As interest in the XAI field grows and new methods and approaches are developed, new survey articles have appeared that attempt to better categorize the approaches in this field. Islam et al. [9] provides an overview of survey articles published in the period from 2017 to 2020. Many of these articles base the proposed taxonomy on categorizing methods on the following criteria:

1) Scope of interpretability - a technique provides an understanding of the entire logic of a model (global) or a specific decision/prediction only (local).
2) Applicability – a technique can be applied to a single type/class of algorithm (model specific) or to any type of AI algorithm (model agnostic).
3) Time of interpretability - model interpretation happens after the model is trained and is independent of its internal design (post hoc) or models are inherently easy to understand (intrinsic).

Table I lists some of the survey papers from 2020 onward. In addition to the general categorization mentioned above, Das and Rad in [10] offer a deeper look at the methodology behind deep learning algorithms by focusing on mathematical summaries of the influential papers. Islam et al. [9] expand the mentioned categorization by showing the application of each of the methods to a common task and analyzing each of them from several perspectives. Tjoa and Guan [14] gave a medical example for each method that they had previously divided into one of two groups of methods that are inherently interpretive (perceptive interpretability) and that ensure interpretability through mathematical structures. Similarly, the authors in the [11] distinguish two groups of XAI approaches. They call them symbolic and sub-symbolic approaches. Examples of symbolic approaches are logical IF-THEN

TABLE I: Review of survey papers from 2020 onward

Survey	Ref.
Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey, 2020	[10]
On the integration of symbolic and sub-symbolic techniques for XAI: A survey, 2020	[11]
Survey of XAI in Digital Pathology, 2020	[12]
A Survey of the State of Explainable AI for Natural Language Processing, 2020	[13]
Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI, 2021	[14]
Explainable Artificial Intelligence Approaches: A Survey, 2021	[9]
Operationalizing Human-Centered Perspectives in Explainable AI, 2021	[15]
Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities, 2021	[16]
Explainable Artificial Intelligence for Tabular Data: A Survey, 2021	[17]
Explainable Artificial Intelligence (XAI) on Time Series Data: A Survey, 2021	[18]
Explainable AI: A Review of Machine Learning Interpretability Methods, 2021	[19]
Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications, 2021	[20]
A survey of visual analytics for Explainable Artificial Intelligence methods, 2022	[21]
A Meta Survey of Quality Evaluation Criteria in Explanation Methods, 2022	[22]

rules. Such an approach is understandable to people but requires the intervention of experts who will manually encode symbolic knowledge in the logic rules. On the other hand, the numerical and statistical representations of the learned model are sub-symbolic, so there is no need for expert intervention, but are not inherently understandable to humans. They recommend and overview the hybrid approaches. The lack of most of the mentioned articles is the overfocus on the algorithm itself. There is a lack of user and stakeholder focus [15]. Also, a small number of XAI approaches are focused on methods that work on data that change its value over time, i.e. time series data. Existing methods mainly provide technical explanations intended for developers. Rojat et al. provide an overview of XAI methods applied on time series data [18]. Apart from the lack of focus on time series data, it is interesting to note that prior to the [17] paper there was no survey paper focusing on tabular data. This form of data is one of the most commonly used today in various sectors such as the financial sector, medicine, education, legal sectors, etc. Despite many existing articles dealing with XAI methods, there is still a lack of an article that offers an overview of specific and commercial applications and the benefits that those applications have brought. There are articles that give examples of application, but mainly relate to a specific domain (e.g. application of XAI within digital pathology [12], NLP [13] or to a specific family of methods [21] or to work with a specific type of data [17].

III. XAI APPLICATIONS

Explainable Artificial Intelligence (XAI) is a core area of AI in which methods are developed to explain the inner logic of either learning algorithms, models that are derived from them, or knowledge-based inference approaches [23].

According to [6], [7] we can distinguish four types of needs for the use of XAI techniques and the possibilities of their application:

A1 Model justification - to explain why a decision was made, especially when an important or unexpected decision is generated, all with the aim of increasing confidence in the model's operation.

A2 Model controlling and debugging – to prevent catastrophic outcomes. A greater understanding of the system increases the visibility of unknown flaws and helps to quickly identify and fix the errors [8].

A3 Model improving - when a user understands why and how a system produced a particular result, he can easily upgrade and improve it and make it smarter, and maybe make it work faster (such as [24], [25] for the development of algorithms). In addition to improving the explanation-generating model, understanding the decisions generated can improve the overall business process based on the decisions of the AI model.

A4 Knowledge discovery - by the appearance of some invisible results of the model and by understanding why and how they arose, one can learn new laws (e.g. discover some new structures in a living organism). Also, since AI agents are often smarter than humans, by understanding their behavior it is possible to learn new skills (e.g. in chess).

A1 Model justification

Sachan et al. [26] proposed an XAI decision-support system that can describe the sequence of events that led to a loan application determination. This system can integrate human knowledge and can through supervised learning learn from previous data. A generated explanation could be delivered in textual form to rejected applicants as grounds for declining their loan applications. In [27], Ohana et al. analyzed the March 2020 stock market crash using gradient-boosted decision trees which the model accurately predicted out of the sample. To explain the generated prediction they used SHAP (SHapley Additive exPlanations). The purpose was to estimate the dangers associated with borrowing credit through peer-to-peer lending systems. Zhong et al. [28] proposed QAjudge, a model based on reinforcement learning whose purpose is to explain legal judgment predictions generated by Legal Judgment Prediction (LJP). QAjudge depicts the forecast generating process in a way that people can understand. This is accomplished by asking human-readable questions repeatedly and then predicting judgments based on the responses [17]. Although deep learning models have proven to be very powerful, the problem is their non-transparency and they are mostly oriented towards image format data (e.g. 2D CNNs). Therefore, the authors [29] propose a method that converts tabular data into images and thus allows the application of 2D CNN in credit scoring. The predictions of the models were then explained by highlighting important pixels that match the

prediction class by using Grad-CAM (Gradient-weighted Class Activation Mapping), LIME (Local Interpretable Model-agnostic Explanations), SHAP values, and Saliency Map method. Therefore, due to various legal regulations and acts it is often required that systems can explain their decisions and that they are not biased in any way ([30], [31]).

A2 Model controlling and debugging

The authors of [32] applied interpretable methods on times series data, precisely ECG data. Their goal was to detect myocardial infarction directly from data. Using the generated explanations, clinicians can verify that the model takes into account relevant specific patterns when making a decision. Ribeiro et al. [33] have shown that using the XAI method can detect data bias and erroneous learning. Namely, the task was to distinguish whether a wolf or a husky was in the picture. They realized using the LIME method that the model learned that the presence of snow in the background in a picture implies a presence of a husky. The authors of [34] also present the detection of an erroneously learned model using the XAI method. Namely, the model learned that the presence of an author's mark in the picture's corner is the main characteristic for the classification that a horse is in the picture. Not much damage has been done in these cases, but the question is what would happen had the stakes been higher.

A3 Model improving

CrystalCandle is an example of a tool that has helped strengthen LinkedIn's business [35]. Introduction of explainations to recommender system for Linkedin software sales team resulted in increased subscription revenue for their software by 8%.

A4 Knowledge discovery

The authors of [36] describe the application of an explainable Deep Learning System for Healthcare using electronic health record (EHR). The model predicts the ultimate risk of heart failure and provides an explanation along with it. Using the LIME method provides insight into which features have a positive effect and which have a negative effect on heart failure. Wu et al. [37], analyzing SARS-COV-2-positive patient data and using interpretive models of machine learning, came to the conclusion which biomarkers indicate severe infection and increased risk of death. Recent medical research has confirmed their conclusions. Although it is hardly possible by an ophthalmologist, DNNs can accurately detect a person's gender from retinal fundus images [38]. Using BagNet, DNN architecture, they found that patches from the optic disc provide predominantly male evidence, whereas patches from the macula provide mostly female evidence [38]. Alpha Zero defeated world chess champion Vladimir Kramnik. Using the sparse linear regression method, it was concluded which AlphaZero figures give the most importance to the various concepts during training. They concluded that in relation to humans who mainly focuses on certain openings, AlphaGo in the early stages tries out various opening movements [39].

Interlude. The new frontiers in explainability are explainable reinforcement learning (XRL) and GNN explainability. The former is challenging because of intricate temporal dynamics which complicates explanations (see [8] for more details). The latter is problematic due to the abstractness of representation for which it is hard to find humanly relatable concepts; graphs are abstract structures that are not completely intuitive to domain experts, let alone laymen.

IV. XAI TECHNIQUES FOR GRAPH NEURAL NETWORKS

With the growing popularity of Graph Neural Networks (GNNs) and their usage in various domains which require explanations for scientific or ethical reasons, such as chemistry and medicine [40], [41], mathematics [42], supply chain optimization [43], or programming language processing [44], a need for GNN XAI methods is rapidly growing. However, the power of these GNN models, especially the more recent ones, is only matched by their complexity and the complexity of the underlying data they work on. Although most, if not all, of the models can be categorized into the (augmented) message-passing [45] paradigm, the more recent ones use increasingly abstract constructs from graph theory and abstract algebra, such as motifs [46], [47] and cell complexes [48], [49]. This alone makes generalized explanation methods very hard to define, but is further made more difficult by the underlying data domains, which can consist of various types of graphs and heterogeneous data. This makes most explanations meaningful only in a specific domain, thus requiring model- and domain-specific explanations. Furthermore, GNN models can differ in the task they try to solve, which are: **node-labeling**, **link-prediction** and **graph-classification**. While a large portion of the models can be adopted for any of the above tasks, defining and generating explanations for different tasks can largely affect how a GNN XAI model is structured. Therefore, some methods like the well known GNNExplainer [50], only work on node-labeling tasks, others such as XGNN [51] only work on graph-classification tasks, however, some models such as SubGraphX [52] can be adopted to any task. A large portion of GNN explanation models focus on the node-labeling task, with a smaller focus on graph-classification, but there is a lack of models which generate explanations for link-prediction tasks.

Different taxonomies have been used in previous surveys. Li et al. [53] use an origin-based taxonomy, where they divide models which adopt previously known XAI methods into the **non-GNN-origin** models, and the **GNN-origin** models which are tailored specifically for GNNs. We focus on the **GNN-origin** methods. Another taxonomy of models is proposed by Yuan et al. [54] which is synonymous to the *scope of interpretability* we have mentioned, where the explanations are either **instance-level (local)** or **model-level (global)**. These are then further decomoposed into subcategories based on the method they use to obtain explanations, where the **local** explanations can be **gradient-**, **decomposition-**, **preturbation-** or **surrogate-**

based, and the **global** are only categorized into the **generation-based**, since they only contain a single model. We provide examples of each of the subcategories, while also mentioning their origin. The **decomposition-based** methods, such as GNN-LRP [55], and **gradient-based** methods, such as Grad-Cam [56] are exclusively from a **non-GNN-origin** and adopt previously known XAI methods for GNNs. The **surrogate-based** method subcategory contains models from both origin-based categories, with GraphLIME [57] being a **non-GNN-origin** method and RelEx [58] being a **GNN-origin** method. These methods try to sample a dataset from the neighborhood of similar graphs and fit a simpler, explainable model on the sampled dataset [54], [58]. These models however, do not directly generate an explanation of the model, but rely on the assumption that the relationships in the neighbouring areas of the input example are simple enough and can be captured by a simpler, more explainable model [54]. However, how these neighbourhoods are defined is unspecified and difficult for graph-structured data. The last of the local subcategories is the **perturbation-based** category which consists of exclusively **GNN-origin** methods and are the most represented in the field of GNN XAI. An example of such a model is the aforementioned GNNExplainer [50]. These methods try to mask out unimportant edges, nodes, or features, thus generating a subgraph consisting of only the important information for the specified prediction [54]. Although Yuan et al. [54] do not include counter-factual models such as [40], [59], we believe they belong into the same category, since they mask out certain edges or nodes, but try to do so minimally while changing the prediction, and therefore, generating minimally different counterexamples. This is specific to the models we have mentioned, however, no other counter-factual models have been proposed within to knowledge. Moving away from instance-level explanation models, **model-level** or **global** explanations, have not been as explored. The only known model is XGNN [51], which generates common subgraphs for a specific class using reinforcement learning. The drawback of the method is, the fact that the model has to be learned for each class we want to explain.

So far we have only mentioned *post-hoc* XAI GNN models which suffer from similar problems as the regular *post-hoc* XAI methods [60], such as bias and misinterpretation. Recently, multiple methods for *intrinsic* explanation learning have been proposed. The SE-GNN [61] model uses a *K-nearest labeled nodes* method, in which it finds the labeled nodes most similar to an unlabeled node that it is trying to classify. The similarity depends on a node's features and the local structure of the labeled node's computation graph. Similarly, ProtGNN [62] compares graphs to a number of learned prototypes in the latent space. The similarity of the graphs is then used in the explanation and prediction process. Although we believe these models offer more reliable explanations than the *post-hoc* methods because of the problems outlined in [60], they are extremely model-specific since the explanations are used in the prediction process, which can limit their use. Also, it is unclear how to categorize these models in the previously used scope-based taxonomy.

There is a lack of standardized metrics and datasets [53] for GNN XAI. The existing experimental works use synthetic or small, domain-specific datasets and evaluation metrics requiring ground-truth data and/or human knowledge [53]. These problems have been addressed in [53], [63], [64] where they support the usage of *faithfulness* metrics, which are different measures that try to assess how well the explanation reflects the reasoning of the model [53] instead of using human knowledge, which can sometimes have pitfalls as outlined in [64]. They describe a set of commonly used metrics and their problems, but also define their own called the *explanation confidence* which addresses the problem of graphs with varying sizes affecting the *sparsity* metric and in turn the *(inverse) fidelity* metric in an unbalanced manner. Another step in the right direction, however, limited to instance-level explanations is given by the authors of [63]. They provide a theoretical analysis of the different GNN XAI model bounds with regards to certain wanted properties, further backed by empirical evaluations. The properties they use are *faithfulness*, *stability* and *(un)fairness preservation*. Most of the analysis they provide focuses on the node-labeling task, but the analysis methods can be extended to other tasks as well. Contrast to the authors in [64], they provide more general metrics for explanation evaluation.

Regarding the real-world applications of GNN XAI, to our knowledge, the only model actually used in a real-world setting has been the xFraud model [65] which generates explanations for a fraud-detection model. Their explanation model is a hybrid explainer that consists of the aforementioned GNNExplainer [50] and edge-centrality weights to quantify the importance of different edges for a prediction. They perform the first quantitative comparison between the GNNExplainer and a human-labeled dataset, which is then additionally used to train the edge-centrality model. They use the model to justify the fraud-prediction model's output to avoid diminishing the customer experience on a trading platform, which we can therefore classify into the **A1** application need.

V. Conclusion

This paper describes the current state of explainable AI. Most of the work is still done in supervised learning and there is a convergence and recycling between the surveys with respect to the main concepts relating to explainability. Much has been done in the last years academically and we feel now is the time to validate the body of research in the real world. There is a lack of real-world usage examples with real impact. There is plenty of digressional abstract work, such as mostly using synthetic datasets in GNN explainability research. Regarding explainability, the frontier of research has moved to explainability of reinforcement learning (XRL) which is mostly unexplored [8] and to explainability of GNNs. Existing XRL work consists of mostly methods inspired and transferred from supervised

learning. There is plenty of opportunity for better, novel, and radically different ideas. GNN XAI methods still lack all-encompassing taxonomies, standardized datasets, and metrics. We can observe the lack of link-prediction and global explanation models, as well as a lack of real-world examples of use and human-evaluation studies.

Furthermore, generally for XAI, there is a lack of standardized datasets and unbiased metrics that makes the models hard to compare, but also formally verify their correctness. So to speak, "they are all winners". Finally, recent results exposing the GNN local explainability bounds [63] and the inherent unfairness in explainability [66] should be seriously taken into account. These phenomena can severely degrade AI safety and security, but they can also be a tool and guide for improvements if cleverly leveraged.

REFERENCES

[1] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *arXiv:1703.00056 [cs, stat]*, Feb. 2017, arXiv: 1703.00056. [Online]. Available: http://arxiv.org/abs/1703.00056

[2] "Amazon scraps secret AI recruiting tool that showed bias against women | Reuters." [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[3] E. Jenn, A. Albore, F. Mamalet, G. Flandin, C. Gabreau, H. Delseny, A. Gauffriau, H. Bonnin, L. Alecu, and J. Pirard, "Identifying challenges to the certification of machine learning for safety critical systems," in *European Congress on Embedded Real Time Systems (ERTS 2020)*, 2020.

[4] "Forhumanity organization." [Online]. Available: https://forhumanity.center/

[5] M. Juric, A. Sandic, and M. Brcic, "AI safety: state of the field through quantitative lens," in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, Sep. 2020, pp. 1254–1259, doi: 10.23919/MIPRO48935.2020.9245153.

[6] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2018, pp. 0210–0215, doi: 10.23919/MIPRO.2018.8400040.

[7] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018, conference Name: IEEE Access.

[8] A. Krajna, M. Brcic, T. Lipic, and J. Doncevic, "Explainability in reinforcement learning: perspective and position," *arXiv preprint arXiv:2203.11547*, 2022, doi: 10.48550/arXiv.2203.11547. [Online]. Available: https://doi.org/10.48550/arXiv.2203.11547

[9] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable Artificial Intelligence Approaches: A Survey," *arXiv:2101.09429 [cs]*, Jan. 2021, arXiv: 2101.09429. [Online]. Available: http://arxiv.org/abs/2101.09429

[10] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," *arXiv:2006.11371 [cs]*, Jun. 2020, arXiv: 2006.11371. [Online]. Available: http://arxiv.org/abs/2006.11371

[11] R. Calegari, G. Ciatto, and A. Omicini, "On the integration of symbolic and sub-symbolic techniques for XAI: A survey," *Intelligenza Artificiale*, vol. 14, pp. 1–25, Sep. 2020.

[12] M. Pocevičiūtė, G. Eilertsen, and C. Lundström, "Survey of XAI in Digital Pathology," in *Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges*, A. Holzinger, R. Goebel, M. Mengel, and H. Müller, Eds. Cham: Springer International Publishing, 2020, pp. 56–88. [Online]. Available: https://doi.org/10.1007/978-3-030-50402-1_4

[13] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A Survey of the State of Explainable AI for Natural Language Processing," *arXiv:2010.00711 [cs]*, Oct. 2020, arXiv: 2010.00711. [Online]. Available: http://arxiv.org/abs/2010.00711

[14] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[15] U. Ehsan, P. Wintersberger, Q. V. Liao, M. Mara, M. Streit, S. Wachter, A. Riener, and M. O. Riedl, "Operationalizing Human-Centered Perspectives in Explainable AI," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, May 2021, no. 94, pp. 1–6. [Online]. Available: https://doi.org/10.1145/3411763.3441342

[16] W. Saeed and C. Omlin, *Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities*, Nov. 2021.

[17] M. Sahakyan, Z. Aung, and T. Rahwan, "Explainable Artificial Intelligence for Tabular Data: A Survey," *IEEE Access*, vol. 9, pp. 135 392–135 422, 2021, conference Name: IEEE Access.

[18] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez, "Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey," *arXiv:2104.00950 [cs]*, Apr. 2021, arXiv: 2104.00950. [Online]. Available: http://arxiv.org/abs/2104.00950

[19] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, p. 18, Jan. 2021, number: 1 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/18

[20] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, Mar. 2021, conference Name: Proceedings of the IEEE.

[21] G. Alicioglu and B. Sun, "A survey of visual analytics for Explainable Artificial Intelligence methods," *Computers & Graphics*, vol. 102, pp. 502–520, Feb. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0097849321001886

[22] H. Löfström, K. Hammar, and U. Johansson, "A Meta Survey of Quality Evaluation Criteria in Explanation Methods," *arXiv:2203.13929 [cs]*, Mar. 2022, arXiv: 2203.13929. [Online]. Available: http://arxiv.org/abs/2203.13929

[23] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf. Fusion*, 2021.

[24] M. Brcic and D. Mlinaric, "Tracking Predictive Gantt Chart for Proactive Rescheduling in Stochastic Resource Constrained Project Scheduling," *Journal of Information and Organizational Sciences*, vol. 42, no. 2, Dec. 2018, number: 2. [Online]. Available: https://www.doi.org/10.31341/jios.42.2.2

[25] M. Brcic, M. Katic, and N. Hlupic, "Planning horizons based proactive rescheduling for stochastic resource-constrained project scheduling problems," *European Journal of Operational Research*, vol. 273, no. 1, pp. 58–66, Feb. 2019. [Online]. Available: https://www.doi.org/10.1016/j.ejor.2018.07.037

[26] S. Sachan, J.-B. Yang, D.-L. Xu, D. E. Benavides, and Y. Li, "An explainable ai decision-support-system to automate loan underwriting," *Expert Systems with Applications*, vol. 144, p. 113100, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419308176

[27] J. J. Ohana, S. Ohana, E. Benhamou, D. Saltiel, and B. Guez, "Explainable ai (xai) models applied to the multi-agent environment of financial markets," in *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer, 2021, pp. 189–207.

[28] H. Zhong, Y. Wang, C. Tu, T. Zhang, Z. Liu, and M. Sun, "Iteratively questioning and answering for interpretable legal judgment prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 1250–1257.

[29] X. Dastile and T. Celik, "Making deep learning-based predictions for credit scoring explainable," *IEEE Access*, vol. 9, pp. 50 426–50 440, 2021.

[30] "Proposal for a REGULATION LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT)," 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[31] B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"," *AI Magazine*, vol. 38, no. 3, pp. 50–57, Oct. 2017, number:

3. [Online]. Available: https://ojs.aaai.org/index.php/aimagazine/article/view/2741

[32] N. Strodthoff and C. Strodthoff, "Detecting and interpreting myocardial infarction using fully convolutional neural networks," *Physiological Measurement*, vol. 40, 11 2018.

[33] 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*, 2016, pp. 1135–1144.

[34] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[35] P. Dave, "AI is explaining itself to humans. And it's paying off," *Reuters*, Apr. 2022. [Online]. Available: https://www.reuters.com/technology/ai-is-explaining-itself-humans-its-paying-off-2022-04-06/

[36] S. Khedkar, P. Gandhi, G. Shinde, and V. Subramanian, "Deep learning and explainable ai in healthcare using ehr," in *Deep learning techniques for biomedical and health informatics*. Springer, 2020, pp. 129–148.

[37] H. Wu, W. Ruan, J. Wang, D. Zheng, B. Liu, Y. Geng, X. Chai, J. Chen, K. Li, S. Li *et al.*, "Interpretable machine learning for covid-19: An empirical study on severity prediction task," *IEEE Transactions on Artificial Intelligence*, 2021.

[38] I. Ilanchezian, D. Kobak, H. Faber, F. Ziemssen, P. Berens, and M. S. Ayhan, "Interpretable gender classification from retinal fundus images using bagnets," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 477–487.

[39] T. McGrath, A. Kapishnikov, N. Tomašev, A. Pearce, D. Hassabis, B. Kim, U. Paquet, and V. Kramnik, "Acquisition of chess knowledge in alphazero," *arXiv preprint arXiv:2111.09259*, 2021.

[40] D. Numeroso and D. Bacciu, "MEG: Generating Molecular Counterfactual Explanations for Deep Graph Networks," *arXiv:2104.08060 [cs]*, Apr. 2021, arXiv: 2104.08060. [Online]. Available: http://arxiv.org/abs/2104.08060

[41] A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, "Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI," *Information Fusion*, vol. 71, pp. 28–37, Jul. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521000142

[42] A. Davies, P. Veličković, L. Buesing, S. Blackwell, D. Zheng, N. Tomašev, R. Tanburn, P. Battaglia, C. Blundell, A. Juhász, M. Lackenby, G. Williamson, D. Hassabis, and P. Kohli, "Advancing mathematics by guiding human intuition with AI," *Nature*, vol. 600, no. 7887, pp. 70–74, Dec. 2021, number: 7887 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41586-021-04086-x

[43] J. Juros, M. Brcic, M. Koncic, and M. Kovac, "Exact solving scheduling problems accelerated by graph neural networks," in *(under review)*, Feb. 2022, doi: 10.13140/RG.2.2.19709.23528/1. [Online]. Available: https://doi.org/10.13140/RG.2.2.19709.23528/1

[44] M. Kovac, M. Brcic, A. Krajna, and D. Krleza, "Towards intelligent compiler optimization," in *under review*, 2022, doi: 10.13140/RG.2.2.29644.49288. [Online]. Available: https://doi.org/10.13140/RG.2.2.29644.49288

[45] P. Veličković, "Message passing all the way up," *arXiv:2202.11097 [cs, stat]*, Feb. 2022, arXiv: 2202.11097. [Online]. Available: http://arxiv.org/abs/2202.11097

[46] X. Chen, R. Cai, Y. Fang, M. Wu, Z. Li, and Z. Hao, "Motif Graph Neural Network," *arXiv:2112.14900 [cs]*, Jan. 2022, arXiv: 2112.14900. [Online]. Available: http://arxiv.org/abs/2112.14900

[47] Z. Yu and H. Gao, "MotifExplainer: a Motif-based Graph Neural Network Explainer," *arXiv:2202.00519 [cs]*, Feb. 2022, arXiv: 2202.00519. [Online]. Available: http://arxiv.org/abs/2202.00519

[48] C. Bodnar, F. Frasca, Y. G. Wang, N. Otter, G. Montúfar, P. Liò, and M. Bronstein, "Weisfeiler and Lehman Go Topological: Message Passing Simplicial Networks," *arXiv:2103.03212 [cs]*, Jun. 2021, arXiv: 2103.03212. [Online]. Available: http://arxiv.org/abs/2103.03212

[49] C. Bodnar, F. Frasca, N. Otter, Y. G. Wang, P. Liò, G. Montúfar, and M. Bronstein, "Weisfeiler and Lehman Go Cellular: CW Networks," *arXiv:2106.12575 [cs, stat]*, Jan. 2022, arXiv: 2106.12575. [Online]. Available: http://arxiv.org/abs/2106.12575

[50] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating Explanations for Graph Neural Networks," *arXiv:1903.03894 [cs, stat]*, Nov. 2019, arXiv: 1903.03894. [Online]. Available: http://arxiv.org/abs/1903.03894

[51] H. Yuan, J. Tang, X. Hu, and S. Ji, "XGNN: Towards Model-Level Explanations of Graph Neural Networks," *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 430–438, Aug. 2020, arXiv: 2006.02587. [Online]. Available: http://arxiv.org/abs/2006.02587

[52] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in Graph Neural Networks: A Taxonomic Survey," *arXiv:2012.15445 [cs]*, Mar. 2021, arXiv: 2012.15445. [Online]. Available: http://arxiv.org/abs/2012.15445

[53] P. Li, Y. Yang, M. Pagnucco, and Y. Song, "Explainability in Graph Neural Networks: An Experimental Survey," *arXiv:2203.09258 [cs]*, Mar. 2022, arXiv: 2203.09258. [Online]. Available: http://arxiv.org/abs/2203.09258

[54] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in Graph Neural Networks: A Taxonomic Survey," *arXiv:2012.15445 [cs]*, Mar. 2021, arXiv: 2012.15445. [Online]. Available: http://arxiv.org/abs/2012.15445

[55] T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K.-R. Müller, and G. Montavon, "Higher-Order Explanations of Graph Neural Networks via Relevant Walks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021, arXiv: 2006.03589. [Online]. Available: http://arxiv.org/abs/2006.03589

[56] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020, arXiv: 1610.02391. [Online]. Available: http://arxiv.org/abs/1610.02391

[57] Q. Huang, M. Yamada, Y. Tian, D. Singh, D. Yin, and Y. Chang, "GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks," *arXiv:2001.06216 [cs, stat]*, Sep. 2020, arXiv: 2001.06216. [Online]. Available: http://arxiv.org/abs/2001.06216

[58] Y. Zhang, D. Defazio, and A. Ramesh, "RelEx: A Model-Agnostic Relational Model Explainer," *arXiv:2006.00305 [cs, stat]*, May 2020, arXiv: 2006.00305. [Online]. Available: http://arxiv.org/abs/2006.00305

[59] A. Lucic, M. ter Hoeve, G. Tolomei, M. de Rijke, and F. Silvestri, "CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks," *arXiv:2102.03322 [cs]*, Feb. 2022, arXiv: 2102.03322. [Online]. Available: http://arxiv.org/abs/2102.03322

[60] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *arXiv:1811.10154 [cs, stat]*, Sep. 2019, arXiv: 1811.10154. [Online]. Available: http://arxiv.org/abs/1811.10154

[61] E. Dai and S. Wang, "Towards Self-Explainable Graph Neural Network," *arXiv:2108.12055 [cs]*, Aug. 2021, arXiv: 2108.12055. [Online]. Available: http://arxiv.org/abs/2108.12055

[62] Z. Zhang, Q. Liu, H. Wang, C. Lu, and C. Lee, "ProtGNN: Towards Self-Explaining Graph Neural Networks," *arXiv:2112.00911 [cs]*, Dec. 2021, arXiv: 2112.00911. [Online]. Available: http://arxiv.org/abs/2112.00911

[63] C. Agarwal, M. Zitnik, and H. Lakkaraju, "Probing GNN Explainers: A Rigorous Theoretical and Empirical Analysis of GNN Explanation Methods," *arXiv:2106.09078 [cs]*, Feb. 2022, arXiv: 2106.09078. [Online]. Available: http://arxiv.org/abs/2106.09078

[64] L. Faber, A. K. Moghaddam, and R. Wattenhofer, "When Comparing to Ground Truth is Wrong: On Evaluating GNN Explanation Methods," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ser. KDD '21. New York, NY, USA: Association for Computing Machinery, Aug. 2021, pp. 332–341. [Online]. Available: https://doi.org/10.1145/3447548.3467283

[65] S. X. Rao, S. Zhang, Z. Han, Z. Zhang, W. Min, Z. Chen, Y. Shan, Y. Zhao, and C. Zhang, "xFraud: Explainable Fraud Transaction Detection," *Proceedings of the VLDB Endowment*, vol. 15, no. 3, pp. 427–436, Nov. 2021, arXiv: 2011.12193. [Online]. Available: http://arxiv.org/abs/2011.12193

[66] M. Brcic and R. V. Yampolskiy, "Impossibility Results in AI: A Survey," *arXiv:2109.00484 [cs]*, Feb. 2022, doi: 10.48550/arXiv.2109.00484. [Online]. Available: https://doi.org/10.48550/arXiv.2109.00484