Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 2 (2025)

The Role of Graph Theory in Explainable AI (XAI)

Dr Preeti Panwar
Associate Professor in Mathematics
Guru Nanak Khalsa College,Karnal

Abstract

These days, deep neural networks (DNNs) are used widely in autonomous vehicles, healthcare, military and other mission-critical systems with direct effect on lives of people. Its use is challenged by the black-box problem of DNN, raising regulatory and ethical concerns like lack of trust. In "Artificial Intelligence (AI)", "Explainable AI (XAI)" promotes a range of techniques, tools, and models to generate high-quality, intuitive, interpretable, and simple explanations of AI decisions.

AI is mostly hyped by the success of machine learning (ML) and deep learning (DL), which is the part of it. AI is related to other areas like "Reasoning and Knowledge Representation" which should be combined to reach the intelligence intended in the 1950s. XAI or Explainable AI refers to the core backup for applying AI in products, especially for industries having critical systems. This study has conducted a review of XAI not just from the perspective of ML, but also from other areas of study like AI Planning. This study highlights challenges related to XAI in the field of AI with a perspective of graph theory and underlying approaches.

Keywords *– Artificial Intelligence, Explainable AI, graph theory, deep learning, machine learning*

1. Introduction

Artificial Intelligence (AI) models, especially the ones using "deep neural networks (DNN)", are turning the way real-world tasks are performed by humans. Recent years have observed a rise in the use of "Machine learning (ML)" models when it comes to automate different aspects of business, science, and social workflow. The rise is partly because of uprise in research in the field of "Deep Learning (DL)" where billions of neuronal parameters have generalized on a specific task. Efficient use of deep learning models in ophthalmology, healthcare, autonomous robots, self-driving vehicles, developmental disorders, audio and speech processing, image processing detection and classification, and cybersecurity indicating the reach of deep learning models (Torres et al, 2018; Lee et al, 2019; Chen et al, 2020; Sayres et al, 2019; Das et al, 2019; Son et al, 2020).

With high-throughput accelerators of AI to improve performance, easier access to powerful compute nodes with cloud computing environments, and access to storage and big-data datasets enables DL providers to test, research, and operate machine learning models in smartphones, small edge devices, and AI-based services for wider exposure using "Application Programming Interfaces (APIs) (Kwasniewska et al, 2019; Zhang et al, 2019). Even from the governments with GDPR laws, recent interest in XAI shows significant realization of trusts, ethics, and bias of AI along with effect of adversarial examples when it comes to fool classifier decisions (Cath et al, 2018; Keskinbora, 2019; Etzioni and Etzioni, 2017; Bostrom and Yudkowsky, 2018; Stahl and Wright, 2018; Weld and Bansal, 2019).

Miller (2019) suggests that curiosity is among the key reasons for people to seek explanations to certain decisions. Another reason is to promote better learning, generate better results, and reiterate model design. Each explanation must be consistent in different data points and generate similar or stable explanation on same point of data (Sokol and Flach, 2020). AI model should be expressive by explanations to improve understanding of mankind, promote impartial decisions, and confidence in making decisions. To maintain trust, transparency, and fairness in decision-making by machine learning, an interpretable solution or explanation is needed for machine learning programs.

The discipline of AI aims to build smart machines to mimic cognitive functions associated with other human minds like problem solving, learning, and addresses AI for systems from different aspects (Russell and Norvig, 2010). From "Machine Learning (ML) to Game Theory, Knowledge Representation and Reasoning (KRR), Robotics, Uncertainty in AI (UAI), Constraint Satisfaction and Search (CSS), Multi-Agent Systems, Computer Vision, Planning and Scheduling, and Natural Language processing (NLP)," all of them are the foundations of AI. All these subfields have specialized, matured, and converged to access the holy grail of AI, "General Artificial Intelligence (GAI)."

Irrespective of rise of innovation based on machine learning based AI programs, explainability is the most concerning question lacking in deep research as in other subfields of AI. However, answering this question of responsibility, explainability, privacy-preserving, validity (like robustness) and trust of AI systems will be connected inherently to the adoption of AI in industry, especially in companies working with critical systems (Figure 1). Explanation could be deciding or debugging smart systems to follow a real-time recommendation to increase user trust and acceptance.

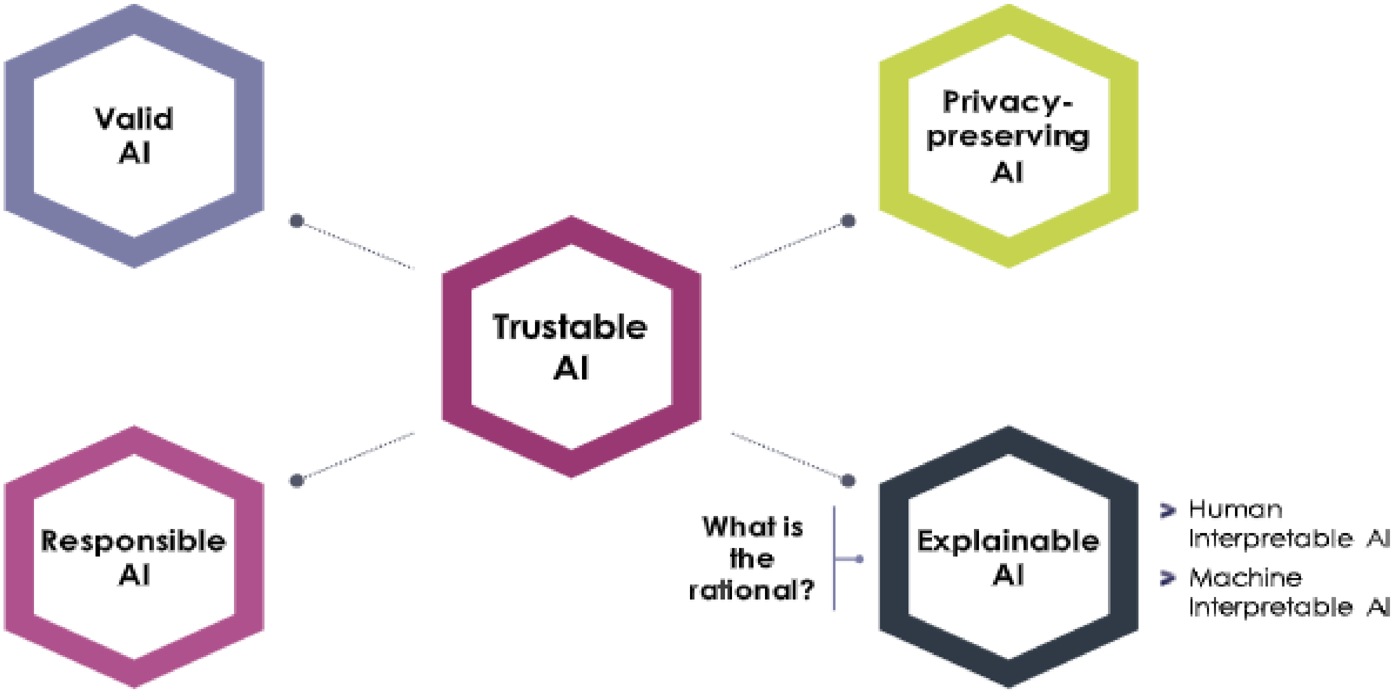

Figure 1 – Components of Trustable AI

Source – Lecue (2020)

With the emergence of most successful AI systems based on machine learning, the same research community will be filling the gap between white-box and black-box ML systems (High, 2012; Silver et al, 2017; Koh and Liang, 2017). Some methods are more successful and AI community is still far from self-explainable AI programs which adapt to any ML model, data, user, context, or application. Works in AI and Web like "Linked Data, Semantic Web, and Knowledge Graphs" are engaged to explain wider family of machine learning programs (Berners-Lee et al, 2023; Bizer et al, 2023; Bollacker et al, 2008). With Knowledge Graphs, the Semantic Web resonates and represents structured data and it must be armed and designed to move XAI closer for understanding of humans (Cudré-Mauroux, 2020).

2. Literature Review

Nandan et al (2025) discussed the complex interplay among graph neural networks (GNN) and Explainable AI (XAI) with huge taxonomy of different explainability approaches for graphical data. The current explainability approaches are classified into self-interpretable and post hoc models. They analyzed practical applications into different fields, showing the importance of transparent GNNs in vital sectors like drug development, fraud detection, and network security. The study also delineates the parameters of evaluation to determine explainability while dealing with common issues in fairness and scalability. The study addresses potential advancements like creating unique XAI approaches designed for GNN architectures, integrating with federated learning, and using these approaches in interdisciplinary areas. The study fills the gap between XAI and GNN with vital resource available for practitioners and researchers to improve efficacy and interpretability of graphical AI solutions.

Knowledge Graphs (KGs) have been applied widely over the years in different areas for various uses. Its semantic model can represent knowledge with a hierarchical structure on the basis of their properties, relationships, and classes of entities. Building large KGs integrates heterogenous data sources and it is helpful to AI programs to be more interpretable and explainable. Rajabi and Etminani (2024) examined recent studies to understand the use of KGs in XAI systems. They designed a framework and split the use of KGs into extracting relationships, features, KG reasoning, and constructing KGs. They also identified where KGs are used most widely in XAI systems (in-model, pre-model, and post-model) as per the above categories. KGs have been used mainly in pre-model XAI for relation and feature extraction. They were used also for reasoning and inference in post-model XAI. There are several studies where KGs are used to explain XAI models in healthcare.

Irrespective of excellent performance and large-scale adoption, ML models are known as "black boxes" as it is not easy to understand their practical operations. In the field of power systems, which needs accountability, it is not easy for experts to justify and trust recommendations and decisions made by such models. In the same way, XAI approaches have been introduced over the past years to improve the explainability of ML models and their output can be understood well. Machlev et al (2022) highlighted the potential of XAI for applications of power systems. Initially, they presented common challenges of XAI in those applications and analyzed recent studies and existing trends in the academia.

DL models have performed well in different industries like finance, healthcare, and autonomous vehicles with advancements in computing technologies and power. Because of black-box DL structures, deciding these learning models should be justified to the end users. XAI explains black box models to reveal underlying decision-making and behavior of models with techniques, algorithms, and tools. Visualization presents prediction explanations and models in an interpretable, explainable way. Alicioglu and Sun

(2022) reviewed existing challenges and trends of visual analytics when it comes to interpret deep learning models with the adoption of XAI approaches and presented future directions of research. They conducted a review of literature on the basis of visual approaches and model usage. They addressed various research questions and discussed research gaps, missing points, and potential directions for future research.

Advanced AI and ML techniques have been very popular over the years because they can solve problems in different areas with high quality and performance. These techniques are very complicated and they often fail to provide understandable and simple explanations for the outputs generated by them. There has been an emergence in the field of XAI. Most data generated are structural in various domains and they include relationships and parts among them. It is possible to represent such data with either simple form or data structure like graph or vector. Haghir Chehreghani (2024) conducted a review on the efficient models to learn from structured data, focusing on how their representation affect the learning models' explainability.

3. Methodology

The study provides detailed insights to different aspects related to graph theory for explainable AI adopted in different industries. We have conducted a systematic review of literature to explore graph theories in major fields of AI. Apart from searching data on databases like Google Scholar, we have considered journals with high-impact factors and peer reviews like IEEE, ACM, Science Direct, and Springer for data analysis for retaining the findings. We have conducted keyword search using terms like explainable AI, XAI, graph theory, machine learning, deep learning, etc."

Inclusion criteria for this study includes (1) Search performed on titles, keywords, and abstracts; (2) studies conducted from 2016 to 2025; (3) studies published in English; and (4) studies published in high-impact factor journals. We have applied a filtering process to choose relevant articles. Exclusion criteria consists of (1) studies not associated to research objectives; (2) studies published on irrelevant publications; and (3) studies not published in English.

4. Graphs for XAI

This study highlights major challenges in the fields of XAI with current approaches, limitations and opportunities for Knowledge Graphs (Figure 2). Areas of AI are broken down with some limitations like natural intersection of domains of AI, questionable limit, and its benefits from well-accepted lists of AI fields, which are represented well in major conferences (Labreuche and Fossier, 2018).

4.1. Machine Learning

ML models use sample data to elaborate a mathematical model, also known as "training data to make decisions or predictions on unseen data or "test data" without being programmed explicitly to perform tasks (Russell and Norvig, 2016). We have studied five key tasks of learning – (1) unsupervised learning to originate some data structured without exposing the labels; (2) supervised learning if data has both labeled and input data; (3) semi-supervised learning with small labeled data; (4) distant learning to use relational data of unlabeled information from current knowledge bases; and (5) reinforcement learning if it is possible to capture further data with interaction.

All tasks of machine learning expose models with abstract or appropriate representation of data. In machine learning, XAI is related to explanation of models and prediction for local explanation. Some models are designed naturally to explicit their linear regression, generalized linear, decision trees, and naïve bayes models. Some of the elements of complex models like partial dependency plot, feature importance, or individual expectation can be useful to capture high representation for global explanation. State-of-the-art approaches revisit feature importance to perform local explanation (Lundberg et al, 2018; Ribeiro et al, 2016).

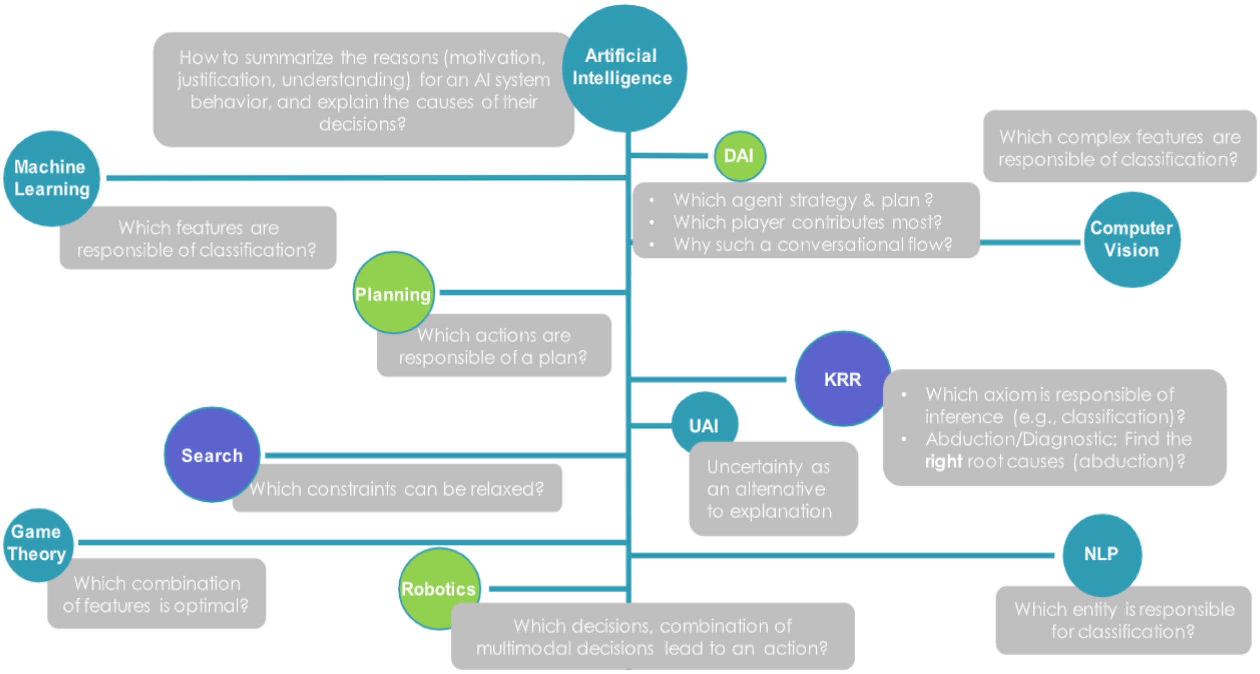
How to summarize the reasons (motivation, justification, understanding) for an AI system behavior, and explain the causes of their decisions?
Artificial Intelligence
Machine Learning
Which features are responsible of classification?
DAI
Which complex features are responsible of classification?
• Which agent strategy & plan ?
• Which player contributes most?
• Why such a conversational flow?
Computer Vision
Planning
Which actions are responsible of a plan?
KRR
• Which axiom is responsible of inference (e.g., classification)?
• Abduction/Diagnostic: Find the right root causes (abduction)?
UAI
Uncertainty as an alternative to explanation
Search
Which constraints can be relaxed?
Game Theory
Which combination of features is optimal?
Robotics
Which decisions, combination of multimodal decisions lead to an action?
NLP
Which entity is responsible for classification?

Figure 2 – Challenges of XAI in major fields of AI (UAI – Uncertainty in AI, DAI – Distributed AI, NLP – Natural Language Processing, KRR – Knowledge Representation and Reasoning)

Source - Lecue (2020)

A lot of methods affect explanation to features in model and data or to counterfactuals or prototypes (Kim et al, 2016; Mittelstadt et al, 2016). Explanation must not be limited to correlation and numerical similarity. Context are encoded by knowledge graphs, while exposing relations and connections, and support causation and inference natively. Current XAI methods in machine learning consider a flat data representation and context is away from the loop of the process of explanation. It is possible to use knowledge graphs for encoding better data representations, defining machine learning model in an interpretable manner and adopting semantic similarity. For instance, knowledge graphs can be limited to input data of ML task to solve some of the different tasks of learning (Han and Sun, 2017). There are also approaches compacting large trees with Knowledge Graphs in random forest or decision trees. For example, it is possible to capture the combinations of nodes as a probabilistic property or concept. When combined, knowledge graphs and ML have a lot of potential from the strength of each other (d’Amato, 2020).

4.2. Artificial Neural Networks

Like any other ML models, ANNs or Artificial Neural Networks look for learning representation. Performance and scalability are the key differentiators with a lot of instances and features, which fit texts and images better. Both global and local explanations are focused strongly on ANN community. Unlike other machine learning models, there is no easy way related to explanation of predictions or ANN models. Current techniques either encode importance of feature with attention mechanism (Ramanishka et al, 2017), attribution (Sundararajan et al, 2017; Shrikumar et al, 2017), or achieve more interpretable estimate with surrogate models (Craven and Shavlik, 1995) like decision tree.

Explanations are built artificially by forcing the nature to focus on some correlations or some groups of features at best. Additionally, there is no representation of any logic of tasks related to learning, making it harder to explain and achieve. ANN seems to work on foundational theory, which derives mathematical model with logical optimizations. Novel architectures of ANN should be designed to encode explanation natively. Some recent models look towards gathering better hierarchical relationships with the model or causality mechanism (Bengio et al, 2019; Hinton et al, 2018). Logical representation layers are added to polish them further in ANN like using the approaches for network dissection, encoding the semantics of outputs, inputs, and their properties (Bau et al, 2017) (Figure 3).

Graph theory could be very vital in a novel design, especially as fresh architectures to embed feature reasoning and embed causation. Makni and Hendler (2019) proposed a layered "graph model representation of graphs in ANN architectures." The layers represent semantics of "Knowledge Graphs" and their predicates and it is captured as 3D adjacency matrices." From the "neural-symbolic reasoning community", other approaches are worth to investigate as ANNs are combined with first order fuzzy logic or probabilistic logic (Donadello et al, 2017; Hitzler et al, 2020; Manheave et al, 2018). The embeddings of knowledge graph are also artifacts of ML where it is possible to elaborate the explanations. ANN could be further advanced by such design by supporting discovery, integration, composition, fragmentation, and reasoning as well.

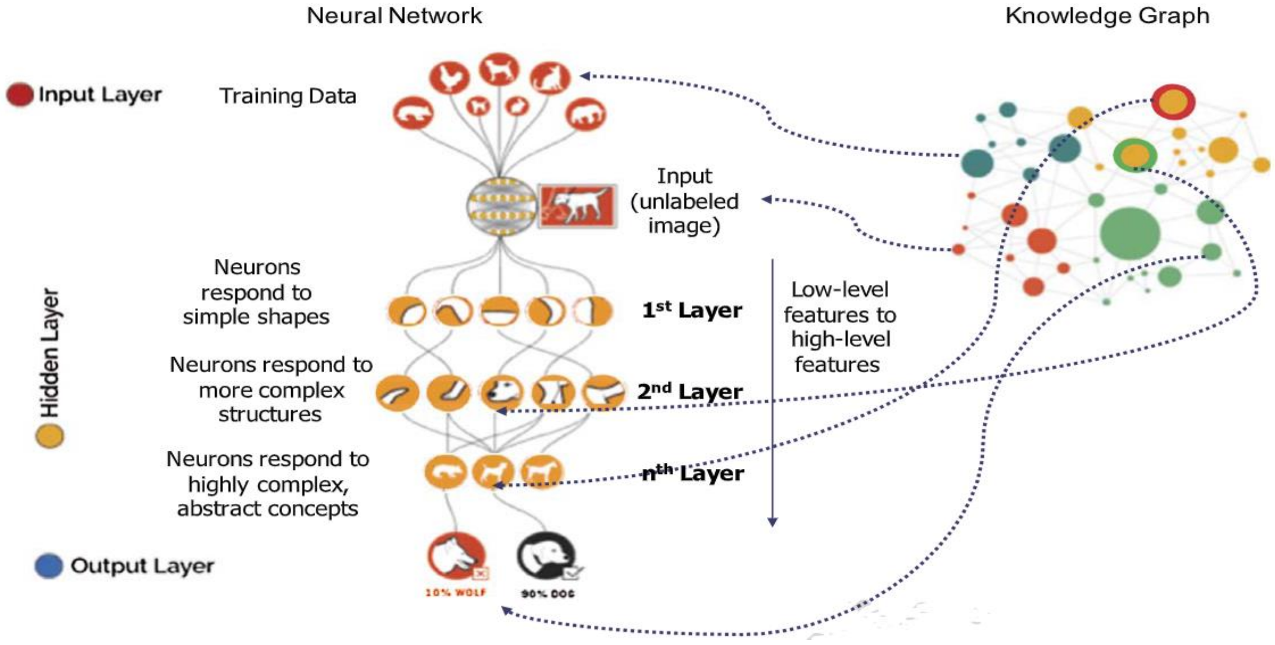
Neural Network
Knowledge Graph
Input Layer
Training Data
Input (unlabeled image)
Hidden Layer
Neurons respond to simple shapes
1st Layer
Low-level features to high-level features
Neurons respond to more complex structures
2nd Layer
Neurons respond to highly complex, abstract concepts
nth Layer
Output Layer
10% WOLF
90% DOG

Figure 3 – Knowledge Graphs giving input/output/training data for XAI neural networks

Source – Zeldam (2018)

4.3. Computer Vision

Computer vision depends upon ANN architectures because of size and nature of data. Tasks related to scene reconstruction, object detection, semantic segmentation, to visual questions are covered in computer vision. Identifying pixels or group of pixels that trigger an uncertainty, an error, or shape detection are the main tasks of XAI in computer vision. Explanation is usually known as visual inspection because of nature of data which is processed. Saliency maps are traditional approaches in Computer Vision (Adebayo et al, 2018). They cover a lot of variants for gathering representative features with gradient modification. Another approach to segment ANN for interpretable layers and units is network dissection (Bau et al, 2017).

Even though interesting artifacts of visualization are exposed by saliency maps, any semantics are not captured. Those artifacts gather disentangled representation at best, which is subject to human interpretation. The semantics of such representation are exposed by Knowledge Graphs. In ANN, semantics are integrated with open challenges of hidden units of space. Semantics can be added with Knowledge Graphs and context to solve open challenges like quantifying factors and detecting them, and disentangled representation.

5. Conclusion

Irrespective of rise in innovation based on XAI systems, industry is facing challenge of applying products at scale, especially for industries working with critical systems. Trust in AI is reveled to coin industry needs to move ahead to the next step. Trustable AI is related to privacy-preserving modeling, responsibility validity, and also explainability. For debugging smart systems, explanation could be used to follow suggestion in real-time, which will increase user trust and acceptance. In AI, explanation has different open challenges, approaches, definitions and meanings, as per which fields of AI is touching the question.

By introducing several solutions, the question has been open in all AI domains. This study has presented challenges of using graph theory in various areas of AI. Significant progress could be achieved only with combinations with semantic layers, such as, explainable AI to empower explanation of complex programs. It is observed that significant research is based on “model-agnostic post-hoc explainability algorithms” because they have great reach and are easy to integrate. In addition, local surrogate models and additive models gathered high interest with super-pixels of information to determine the attributions of input feature.

Studies have discussed limitations of visualizations of explanation maps as there is a shift from gradient-based and perturbation models because of input invariances and adversarial attacks. It is still challenging to evaluate those methods and they pose open research question in XAI research. Field of CAI is still growing and XAI approaches must be selected carefully and developed.

References

- Torres, A. D., Yan, H., Aboutalebi, A. H., Das, A., Duan, L., & Rad, P. (2018). Patient facial emotion recognition and sentiment analysis using secure cloud with hardware acceleration. In *Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications* (pp. 61-89). Academic Press.

- Lee, S. M., Seo, J. B., Yun, J., Cho, Y. H., Vogel-Claussen, J., Schiebler, M. L., ... & Kim, N. (2019). Deep learning applications in chest radiography and computed tomography: current state of the art. *Journal of thoracic imaging*, *34*(2), 75-85.
- Chen, R., Yang, L., Goodison, S., & Sun, Y. (2020). Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. *Bioinformatics*, *36*(5), 1476-1483.
- Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., ... & Webster, D. R. (2019). Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*, *126*(4), 552-564.
- Das, A., Rad, P., Choo, K. K. R., Nouhi, B., Lish, J., & Martel, J. (2019). Distributed machine learning cloud teleophthalmology IoT for predicting AMD disease progression. *Future Generation Computer Systems*, *93*, 486-498.
- Son, J., Shin, J. Y., Kim, H. D., Jung, K. H., Park, K. H., & Park, S. J. (2020). Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology*, *127*(1), 85-94.
- Kwasniewska, A., Szankin, M., Ozga, M., Wolfe, J., Das, A., Zajac, A., ... & Rad, P. (2019, October). Deep learning optimization for edge devices: Analysis of training quantization parameters. In *IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society* (Vol. 1, pp. 96-101). IEEE.
- Zhang, C., Patras, P., & Haddadi, H. (2019). Deep learning in mobile and wireless networking: A survey. *IEEE Communications surveys & tutorials*, *21*(3), 2224-2287.
- Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the ‘good society’: the US, EU, and UK approach. *Science and engineering ethics*, *24*, 505-528.
- Keskinbora, K. H. (2019). Medical ethics considerations on artificial intelligence. *Journal of clinical neuroscience*, *64*, 277-282.
- Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, *21*, 403-418.
- Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In *Artificial intelligence safety and security* (pp. 57-69). Chapman and Hall/CRC.
- Stahl, B. C., & Wright, D. (2018). Ethics and privacy in AI and big data: Implementing responsible research and innovation. *IEEE Security & Privacy*, *16*(3), 26-33.
- Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Communications of the ACM*, *62*(6), 70-79.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.
- Sokol, K., & Flach, P. (2020, January). Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 56-67).
- Russell, S., & Norvig, P. (2010). Artificial intelligence: a modern approach. 3rd. *Upper Saddle River, EUA: Prentice-Hall*.

- Lecue, F. (2020). On the role of knowledge graphs in explainable AI. *Semantic Web, 11*(1), 41-51.
- High, R. (2012). The era of cognitive systems: An inside look at IBM Watson and how it works. *IBM Corporation, Redbooks, 1*, 16.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. (2017). Mastering the game of go without human knowledge. *nature, 550*(7676), 354-359.
- Koh, P. W., & Liang, P. (2017, July). Understanding black-box predictions via influence functions. In *International conference on machine learning* (pp. 1885-1894). PMLR.
- Berners-Lee, T., Hendler, J., & Lassila, O. (2023). The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. In *Linking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web* (pp. 91-103).
- Bizer, C., Heath, T., & Berners-Lee, T. (2023). Linked data-the story so far. In *Linking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web* (pp. 115-143).
- Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008, June). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 1247-1250).
- Cudré-Mauroux, P. (2020). Leveraging knowledge graphs for big data integration: the XI pipeline. *Semantic Web, 11*(1), 13-17.
- Nandan, M., Mitra, S., & De, D. (2025). GraphXAI: a survey of graph neural networks (GNNs) for explainable AI (XAI). *Neural Computing and Applications*, 1-52.
- Rajabi, E., & Etminani, K. (2024). Knowledge-graph-based explainable AI: A systematic review. *Journal of information science*, *50*(4), 1019-1029.
- Machlev, R., Heistrene, L., Perl, M., Levy, K. Y., Belikov, J., Mannor, S., & Levron, Y. (2022). Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, *9*, 100169.
- Alicioglu, G., & Sun, B. (2022). A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics*, *102*, 502-520.
- Haghir Chehreghani, M. (2024). A review on the impact of data representation on model explainability. *ACM Computing Surveys*, *56*(10), 1-21.
- Labreuche, C., & Fossier, S. (2018, July). Explaining Multi-Criteria Decision Aiding Models with an Extended Shapley Value. In *IJCAI* (pp. 331-339).
- Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. pearson.
- Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " 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* (pp. 1135-1144).
- Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, *29*.
- Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279-288).
- Han, X., & Sun, L. (2017, February). Distant supervision via prototype-based global representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).
- d'Amato, C. (2020). Machine learning for the semantic web: Lessons learnt and next research directions. *Semantic Web*, *11*(1), 195-203.

- Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319-3328). PMLR.
- Shrikumar, A., Greenside, P., & Kundaje, A. (2017, July). Learning important features through propagating activation differences. In *International conference on machine learning* (pp. 3145-3153). PMlR.
- Ramanishka, V., Das, A., Zhang, J., & Saenko, K. (2017). Top-down visual saliency guided by captions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7206-7215).
- Craven, M., & Shavlik, J. (1995). Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, *8*.
- Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., ... & Pal, C. (2019). A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*.
- Hinton, G. E., Sabour, S., & Frosst, N. (2018, February). Matrix capsules with EM routing. In *International conference on learning representations*.
- Makni, B., & Hendler, J. (2019). Deep learning for noise-tolerant RDFS reasoning. *Semantic Web*, *10*(5), 823-862.
- Donadello, I., Serafini, L., & Garcez, A. D. A. (2017). Logic tensor networks for semantic image interpretation. *arXiv preprint arXiv:1705.08968*.
- Hitzler, P., Bianchi, F., Ebrahimi, M., & Sarker, M. K. (2020). Neural-symbolic integration and the semantic web. *Semantic Web*, *11*(1), 3-11.
- Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., & De Raedt, L. (2018). Deepproblog: Neural probabilistic logic programming. *Advances in neural information processing systems*, *31*.
- Zeldam, S. G. (2018). *Automated failure diagnosis in aviation maintenance using explainable artificial intelligence (XAI)* (Master's thesis, University of Twente).
- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, *31*.
- Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6541-6549).