International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Explainable Artificial Intelligence (XAI): Shedding Light on AI's Black Box

Kinjal Gandhi[1], Nihali Jain.[2], Milind Shah[3]*, Premal Patel[4], Neeta Chudasama[5], A.Vani Lavanya[6]

Submitted: 27/01/2024 **Revised**: 05/03/2024 **Accepted**: 13/03/2024

Abstract: The rise of AI, especially in critical domains, has raised transparency and accountability concerns due to opaque black-box algorithms. This article explores Explainable AI (XAI) and its application, focusing on Remote Sensing and Signal Processing. AI is increasingly used in sectors like autonomous driving, healthcare, and finance, necessitating transparent decision-making. Opaque AI models impact trust, bias, and accountability, driving the need for XAI. XAI provides insights into AI decision rationale, supported by GDPR's right to explanations. In Reinforcement Learning (RL), XAI faces unique challenges due to RL's sequential nature and the lack of human-labeled data. XAI methods include model interpretation, post-hoc explanations, interactive explanations, and hybrid approaches. XAI categories include transparent models, opaque models, model-agnostic and model-specific approaches, explanation by simplification, explanation by feature relevance, visual explanation, and local explanation. Applications span healthcare, criminal justice, natural language processing, autonomous systems, agriculture, finance, computer vision, forecasting, remote sensing, social media, and transportation, enhancing trust and fairness. Challenges in natural language generation involve evaluating, handling ambiguous language, constructing narratives, and communicating data quality. This article highlights XAI's role in making AI transparent, addressing black-box algorithm challenges, and fostering trust and accountability in AI decision-making.

Keywords: *Artificial Intelligence, Natural Language Generation, Explainable Artificial Intelligence.*

1. Introduction

Explainable Artificial Intelligence (XAI) is emerging as a pivotal domain within the broader field of Artificial Intelligence (AI), addressing the inherent opacity and complexity of AI systems. As AI technologies are increasingly integrated into sensitive domains with significant societal and ethical implications, transparency, trust, and accountability are paramount. Applications ranging from autonomous driving to medical diagnosis and business optimization underscore the urgency to decipher the decision-making processes of AI systems [1]. While AI, especially Machine Learning (ML), has showcased remarkable capabilities, the inner workings of its black-box algorithms remain elusive to both end-users and even some data scientists.

The opacity of AI systems raises questions of trust, potential bias, accountability, and comprehensibility. The challenge arises because AI operates like a black box, especially Machine Learning. Input data goes through a neural network, yielding outputs without revealing the intermediary step. The skepticism of transparency generates skepticism when the AI produces unexpected outcomes. Such mistrust can lead to rejecting AI's conclusions, hampering its potential benefits.

Addressing this issue, Explainable AI (XAI) seeks to illuminate the inner workings of AI algorithms, enabling explanations for their decisions. XAI aims to foster trust, debug biases, and enhance accountability by unveiling the rationale behind AI's actions. Because the General Data Protection Regulation of the European Union requires that individuals have the right to explanations, the significance of XAI is further highlighted [2].

XAI particularly gains significance in Reinforcement Learning (RL), a branch of AI where agents learn optimal actions through interactions with an environment. RL's

[1] *Department of Computer Engineering, Devang Patel Institute of Advance Technology & Research (DEPSTAR), Faculty of Technology & Engineering (FTE), Charotar University of Science & Technology (CHARUSAT), Changa, Gujarat, India*
Email: kinjal445@gmail.com
ORCID – 0000-0002-6791-075X
[2] *School of Pharmacy, ITM (SLS) Baroda University, Paldi, Near Jarod, Vadodara, Gujarat, India. 391510*
Email: jainnihali@gmail.com
ORCID – 0009-0001-9549-866X
[3]*Department of Computer Engineering, Devang Patel Institute of Advance Technology & Research (DEPSTAR), Faculty of Technology & Engineering (FTE), Charotar University of Science & Technology (CHARUSAT), Changa, Gujarat, India*
Email: milindshahcomputer@gmail.com
ORCID – 0009-0001-6077-3924
[4] *Department of Computer Engineering, Devang Patel Institute of Advance Technology & Research (DEPSTAR), Faculty of Technology & Engineering (FTE), Charotar University of Science & Technology (CHARUSAT), Changa, Gujarat, India*
Email: premalj_patel@yahoo.com
ORCID – 0009-0009-8521-633X
[5]*Department of Computer Engineering, Devang Patel Institute of Advance Technology & Research (DEPSTAR), Faculty of Technology & Engineering (FTE), Charotar University of Science & Technology (CHARUSAT), Changa, Gujarat, India*
Email: neeta.chudasama2011@gmail.com
ORCID – 0009-0005-1373-8973
[6]*Assistant Professor, Department of Computer Science & Engineering, St. Joseph's Institute of Technology, Chennai, Tamil Nadu, India*
Email: vanilavanya8@gmail.com
ORCID – 0000-0001-7136-0511
** Corresponding Author Email: milindshahcomputer@gmail.com*

dynamic nature and long sequences of actions intensify the need for understandable explanations. Traditional AI models often focus on text-style explanations or visualizations, aiding human-understandable interpretation. [3,4] Visualizations, such as "saliency maps," offer insights into critical areas of input images influencing outcomes [5]. While previous reviews have explored XAI in broader ML contexts, this work zooms in on RL due to its unique challenges and potential advantages. RL's inherent sequential nature necessitates explanations that encompass sets of interrelated actions. The lack of human-labeled training data in RL poses challenges in generating human-readable explanations. Moreover, the corporate interests of maintaining proprietary information and competitiveness can hinder the development of transparent AI systems [6].

In the realm of providing eXplainable Artificial Intelligence (XAI) for Reinforcement Learning (RL), notable challenges arise that stem from the nature of RL itself. One prominent challenge pertains to the inherent complexity of RL systems. Unlike conventional Machine Learning (ML) techniques, where decisions are often isolated or unrelated, RL involves a sequence of interconnected choices made over time. This sequential decision-making, often in real-time, demands explanations encompassing a coherent set of actions connected by a purposeful trajectory.

Another challenge emerges from the training paradigm of RL agents. Typically, RL agents acquire knowledge through interaction with their environment rather than being trained on explicit datasets. This process relies on feedback loops generated by environmental observations and actions. Consequently, crafting human-readable explanations becomes intricate. While the spaces of observations and actions can be well-defined, the absence of human-labeled training data that explicitly links actions and observations complicates the task of generating meaningful and coherent explanations [7]. Additionally, the pursuit of developing transparent and explainable AI systems encounters further complexities. The commercial interests of companies can potentially clash with the goals of explainability. Striking a balance between transparency and safeguarding proprietary information becomes a delicate task. Companies may be reluctant to expose intricate details that could reveal strategic interests or sensitive information. Furthermore, implementing XAI can entail additional costs in terms of development efforts and potential trade-offs with competitive advantage.

2. RELATED WORK

Reference	Relevant Content	Literature Survey
Glass, A. et al.	Explores trust in adaptive agents for XAI.	Trust is crucial in the context of adaptive agents. Transparency and a clear understanding of agent behavior contribute to establishing trust in Explainable AI.
Anjomshoae, S. et al.	Systematic review on XAI in autonomous agents.	The purpose of this comprehensive literature review is to shed light on the ever-changing environment of explainable artificial intelligence, with particular focus on its applications in multiagent systems and autonomous agents.
Adadi, A. & Berrada, M.	Survey on XAI, peeking inside black-box models.	Adadi and Berrada's survey delves into the intricacies of Explainable AI, providing insights into methods for unveiling the black-box nature of machine learning models.
Gilpin, L. H. et al.	Overview of machine learning interpretability.	The paper emphasizes the significance of interpretability, shedding light on the 'why' behind machine learning model decisions.
Rudin, C.	Advocates for interpretable models.	Rudin's work underscores the urgency of adopting interpretable models, particularly in contexts where high-stakes decisions are made, challenging the use of black-box models.

Ali, S. et al.	XAI for achieving Trustworthy AI.	The paper navigates the landscape of Trustworthy AI, highlighting the role of Explainable AI in building and maintaining trust in artificial intelligence systems.
Amann, J. et al.	Multidisciplinary perspective on XAI in healthcare.	Amann et al.'s work delves into the intricacies of explainability in healthcare AI, addressing the multidisciplinary challenges and emphasizing the need for trust in medical decision-making.
Adebayo, J. et al.	Sanity checks for saliency maps.	Adebayo et al.'s work focuses on sanity checks for saliency maps, ensuring their robustness and reliability in interpreting deep learning models.
Mohanty, S. & Vyas, S.	Strategies for businesses in the age of AI.	Mohanty and Vyas provide strategic insights for businesses, advocating the adoption of collaborative human-machine strategies to thrive in the age of artificial intelligence.
Baker, B. et al.	Emergent tool use from multi-agent auto curricula.	Baker et al. delve into the fascinating realm of emergent tool use, highlighting the autonomous development of tools in multi-agent systems.
Pasquale, F.	Exploration of secretive algorithms.	Pasquale's work sheds light on the secretive nature of algorithms, unraveling the control they exert over money and information in society.
Carey, P.	Legal aspects of data protection.	Carey explores the legal dimensions of data protection, offering a practical guide to the complex landscape of UK and EU laws governing data handling.

3. Natural Language Generation

When writing explanations in natural language, it is important to consider clarity, utilitarianism, and accessibility. To ensure high-quality explanations, they should be customized for specific objectives and audiences, have a narrative structure, and address uncertainty and data reliability [56]. According to, there are four major obstacles to overcome to develop high-quality explanations:

1. Evaluation Challenge: There is a need for reliable and cost-effective techniques to assess the quality of explanations at different levels of rigor, such as scrutability and trust. Automated evaluation measures for natural language creation have been classified and revised to address this challenge [57].

2. Challenge of Ambiguous Language: Although qualitative, Ambiguous language can improve human comprehension. However, ensuring that users understand ambiguous language correctly and preventing misconceptions can be difficult. Ranking messages based on user comfort with features and concepts can help address this challenge. Additionally, using natural language and appropriate terminology is essential [58].

3. Narrative Challenge: Teaching symbolic reasoning using stories instead of statistics and probabilities is more understandable. Developing algorithms for constructing narrative justifications is crucial in addressing this challenge [59].

4. Challenge of Communicating Data Quality: Techniques are needed to alert consumers when data problems affect the results. The dataset used in the AI system's development will determine the justifications offered for its output and findings. Data quality issues can be harmful, including bias, incompleteness, and inaccuracy. The use of poor data during

the training of AI systems can be seen in the outcomes. For example, given the differences in their polluted surroundings, an AI system created for forecasting lung cancer risks based on American data might not produce an accurate risk estimate for a resident of a South Asian country [60].

Furthermore, in the specific context of AI systems producing lengthy textual reports in medical fields, ensuring that the generated reports resemble doctors' behavior and are coherent is difficult. Transformer networks, such as language model decoders, can be used to maintain word relationships in longer sentences. Evaluating these generated reports requires comparisons with reports produced by humans. Still, removing unnecessary information from human-generated reports before contrast is important, as they are often free-text and not bound by templates [61].

4. Taxonomy of eXplainable Artificial Intelligence (XAI)

Within the literature, a range of terms exist to address the challenge posed by the "black box" nature of certain AI, ML, and DL models. The following distinctions are notable:

• Transparency: Transparency signifies a model's inherent potential for comprehensibility. In essence, it directly contrasts the concept of a "black box" [8].

• Interpretability: This pertains to the ability to provide human-readable explanations. The aim is to make complex model decisions understandable for humans [9].

• Explainability: This involves creating a bridge between humans and AI systems through explanations. It encompasses AI systems that are accurate and easily grasped by humans [9]. Explainability is a heavily debated topic with far-reaching implications that extend beyond the technical properties of AI. Even though research indicates that AI algorithms can outperform humans in certain analytical tasks (e.g., pattern recognition in imaging), the lack of explainability has been criticized in the medical domain [12].

To contribute to the discourse on explainable AI in medicine, this paper draws attention to the interdisciplinary nature of explainability and its implications for the future of healthcare.

There are several methods for achieving XAI, including [11]

1. Model interpretation involves analyzing the internal workings of an AI model to understand how it makes decisions.

2. Post-hoc explanation: Explaining how the AI model has decided.

3. Interactive explanation: This involves allowing humans to interact with the AI model to understand how it makes decisions.

4. Hybrid explanation: This involves combining multiple methods to achieve XAI.

While these terms share semantic similarities, they delineate different levels of AI's acceptability to humans. A more detailed ontology and taxonomy of eXplainable AI (XAI) at a broader level can be outlined as follows:

• Transparent Model: Typical transparent models include k-nearest neighbors (kNN), decision trees, rule-based learning, Bayesian networks, and similar models. The decisions from these models are usually transparent, but transparency alone doesn't guarantee ready explainability [13].

• Opaque Model: Opaque models encompass random forests, neural networks, SVMs, etc. Despite high accuracy, these models lack transparency [14].

• Model Agnostic: Model-agnostic XAI approaches are designed for broad applicability. They remain adaptable without depending on a model's inherent architecture, functioning by relating input to output [15].

• Model-Specific: Model-specific XAI approaches target specific models to bring transparency to particular types.

• Explanation by Simplification: This approach simplifies a model via approximation, creating alternative models to elucidate predictions. For instance, a linear model or decision tree could be built around complex predictions for explanation [17].

Explanation by Feature Relevance: This approach evaluates a feature based on its contribution to a model's decision, considering all possible combinations [18] [37].

• Visual Explanation: This type of XAI leverages visualization for interpreting predictions or decisions from input data [19].

• Local Explanation: Local explanations approximate a model within a specific area, offering insight into its operation for similar inputs [20].

The ML literature predominantly employs "interpretability" over "explainability." However, assert that interpretability does not address all issues with understanding "black-box" models. Explainability is crucial to garnering user trust and meaningful insights into such approaches' causes, rationales, and decisions. While explainable models are inherently interpretable, the reverse is not universally true [21]. The existing literature categorizes XAI taxonomy based on:

1. Scope (local and global)- XAI can be categorized based on the scope of explanation, which can be local or global.

Local Explanations: Local explanations focus on providing insights into specific model predictions. They answer questions like, "Why did the AI classify this image as a cat?" Local explanations are valuable for understanding individual model decisions.

Global Explanations: Global explanations, on the other hand, aim to provide a holistic view of how an AI model operates. They answer questions like, "What are the key features that influence this model's overall behavior?" Global explanations help comprehend the model's functioning as a whole [22, 62, 63, 64].

2. Usage (post hoc and intrinsic to model architecture) - Another way to categorize XAI is based on when and how explanations are generated, which can be post hoc or intrinsic to model architecture.

Post Hoc Explanations: Post hoc explanations are generated after the AI model has been decided. They are like retroactive insights into why a specific outcome occurred. Techniques like LIME (Local Interpretable Model-Agnostic Explanations) fall into this category. Post hoc explanations are valuable for understanding the "black box" after the fact.

Intrinsic Explanations: Intrinsic explanations are integrated into the model's architecture and are generated simultaneously with predictions. Intrinsic XAI methods aim to build models that inherently provide explanations as they operate. While these methods are relatively newer and more challenging to develop, they promise real-time interpretability. [65, 66, 67]

3. Methodology (focused on features or model parameters) - Lastly, XAI can be categorized based on the methodology employed, which can focus on features or model parameters.

Feature-Based Explanations: Feature-based XAI methods explain the role and importance of input features. They aim to identify which input features influenced a particular prediction. Feature attribution methods like SHAP (Shapley Additive exPlanations) fall into this category. These methods are widely used to understand image classification, natural language processing, and more.

Model Parameter-Based Explanations: Model parameter-based explanations delve into the inner workings of the AI model itself. They focus on how the model's parameters and internal structures contribute to its decisions. Understanding these aspects can be crucial for experts fine-tuning models or diagnosing issues [22, 68].

Acknowledging the escalating significance of this topic, NIST published Four Principles of XAI in August 2020. These principles define the fundamental characteristics that an AI system must uphold to qualify as XAI: [23]

1. Explanation: The AI system must furnish evidence, support, or reasoning for each decision.

2. Meaningful: The explanation must be intelligible and pertinent to users, catering to user groups' diverse characteristics and requirements.

3. Accuracy: The explanation must precisely reflect the system's processes.

4. Knowledge Limits: The AI system must identify cases beyond its design scope, whose answers may be unreliable.

NIST's publication of the Four Principles of XAI has profoundly impacted the AI community. It has been a guideline for researchers, developers, and policymakers working to make AI systems more transparent and accountable.

Table 1: Methods for eXplainable Artificial Intelligence (XAI)

Name of the method	Description	References
Features-Oriented Methods		
Shapley Additive exPlanation (SHAP)	A game-theoretic approach representing features as players in a coalition game. Computes Shapley values to measure feature contributions. Allows consistent local and global interpretations.	[24]
Class Activation Maps (CAMs)	Applied to CNNs. Represents the per-class weighted linear sum of visual patterns. Highlights influential areas in images through heatmap representation.	[25]
Gradient-weighted Class Activation Mapping (Grad-CAM)	Generalizes CAM to arbitrary CNN architectures without retraining. Computes importance score based on gradients. Produces coarse-grained visualizations.	[26] [27]
Global Methods		
Global Attribution	Explains neural network predictions across subpopulations.	[28]

Mappings (GAMs)	Utilizes rank distance matrix and clustering algorithm to group local feature importances into clusters.	
Gradient-based Saliency Maps	Visualizes absolute gradient values of the majority predicted class. Highlights influential areas in images.	[29]
Deep Attribute Maps	Multiplies output gradient with respective input for heatmap explanation. Indicates positive and negative contributions to output decisions. Sensitive to noisy gradients and input variations.	[30]
Concept Models		
Concept Activation Vectors (CAVs)	Maps human understandable features to neural network's latent features. Represents the degree of abstract features pointing towards chosen concepts.	[31]
Automatic Concept-based Explanations	Extracts CAVs automatically without human bias.	[32]
Surrogate Models		
Local Interpretable Model-Agnostic Explanations (LIME)	Trains interpretable surrogate model to explain global "black box" model predictions. Divides input image into patches for local model training.	[33]
Local, Pixel-based Methods		
Layer-wise Relevance Propagation (LRP)	Uses predefined rules to explain multilayered neural network's output with heatmap. Highlights pixels contributing to the model's prediction.	[34]
DeconvNet	Utilizes semantic segmentation to learn deconvolution network and provide pixel contribution insights during classification.	[35]
Human-Centric Methods		
Human-Centric Approach	Considers explainability as a human-centric phenomenon. Focuses on human reasoning, associations, and analogies, unlike statistical methods.	[36]

These methods span various categories and approaches to enhance the explainability of AI models. They offer insights into feature contributions, pixel influence, and global attributions. However, they differ in their effectiveness and ability to provide human-understandable explanations. A human-centric approach bridges the gap between AI and human reasoning, emphasizing similarity and associations.

5. Conclusion

In conclusion, the evolution of Explainable Artificial Intelligence (XAI) emerges as a critical response to the inherent opacity of AI systems, particularly in domains with significant societal impact. Despite the hurdles presented by RL's complex nature and companies' intrinsic reluctance to disclose proprietary information, the article highlights the importance of balancing transparency and competitive interests. Additionally, the discussion extends to the challenges and considerations in natural language generation for AI explanations, emphasizing the importance of clarity, utility, and accessibility. Overall, the integration of XAI addresses the challenges of black-box algorithms and lays the foundation for responsible and trustworthy AI systems in an increasingly interconnected world.

Author contributions

Kinjal Gandhi, A. Vani Lavanya: Conceptualization, Introduction to XAI, Existing Research Limitations **Nihali Jain:** Conceptualization, Data curation, Writing-Original draft preparation **Milind Shah:** Taxonomy of XAI, Literature Survey, Writing-Reviewing and Editing. **Premal Patel, Neeta Chudasama:** Natural Language Generation, Introduction.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Glass, A., McGuinness, D. L., & Wolverton, M. (2008). Toward establishing trust in adaptive agents. In Proceedings of the 13th International Conference on

Intelligent User Interfaces (pp. 227–236). New York, NY. doi: 10.1145/1378773.1378804.

[2] Carey, P. (2018). Data Protection: A Practical Guide to UK and EU Law. Oxford University Press, Inc.

[3] Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: results from a systematic literature review. In 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019. International Foundation for Autonomous Agents and Multiagent Systems (pp. 1078–1088).

[4] Explainable and Transparent AI and Multi-Agent Systems: Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected PapersTowards Explainable Practical Agency: A Logical Perspective.

[5] Adebayo, J., Gilmer, J., Mueller, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. arXiv [Preprint] arXiv:1810.03292.

[6] Mohanty, S., & Vyas, S. (2018). How to Compete in the Age of Artificial Intelligence: Implementing a Collaborative Human-Machine Strategy for Your Business. Apress. doi: 10.1007/978-1-4842-3808-0.

[7] Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., et al. (2019). Emergent tool use from multi-agent autocurricula. arXiv [Preprint] arXiv:1909.07528.

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

[9] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE (pp. 80–89).

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

[11] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805.

[12] Amann, J., Blasimme, A., Vayena, E., et al. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20, 310. https://doi.org/10.1186/s12911-020-01332-6.

[13] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.

[14] Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061.

[15] Dieber, J., & Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME. arXiv preprint arXiv:2012.00093.

[16] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One, 10, e0130140.

[17] Tritscher, J., Ring, M., Schölr, D., Hettinger, L., & Hotho, A. (2020). Evaluation of post-hoc XAI approaches through synthetic tabular data. In International Symposium on Methodologies for Intelligent Systems (pp. 422–430). Springer.

[18] Chen, H., Lundberg, S., & Lee, S.-I. (2019). Explaining models by propagating Shapley values of local components. arXiv preprint arXiv:1911.11888.

[19] Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter conference on applications of computer vision (WACV) (pp. 839–847)

[20] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626).

[21] Burkart, N., & Huber, M. F. (2020). A survey on the explainability of supervised machine learning. arXiv preprint arXiv:2011.07876.

[22] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.

[23] Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2020). Four principles of explainable artificial intelligence.

[24] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[25] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization.

[26] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618–626).

[27] Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter conference on applications of computer vision (WACV) (pp. 839–847).

[28] Ibrahim, M., Louie, M., Modarres, C., & Paisley, J. (2019). Global explanations of neural networks: Mapping the landscape of predictions. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19 (pp. 279–287). https://doi.org/10.1145/3306618.3314230.

[29] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualizing image classification models and saliency maps. arXiv:1312.6034 [cs].

[30] Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. http://arxiv.org/abs/1711.06104.

[31] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2021). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). http://arxiv.org/abs/1711.11279.

[32] Ghorbani, A., Wexler, J., Zou, J., & Kim, B. (2019). Towards automatic concept-based explanations. http://arxiv.org/abs/1902.03129.

[33] Dieber, J., & Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME. arXiv preprint arXiv:2012.00093.

[34] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One, 10, e0130140.

[35] Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE international conference on computer vision (pp. 1520–1528).

[36] Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). Neural Networks, 130, 185–194.

[37] Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box AI decision systems. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, pp. 9780–9784).

[38] Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923.

[39] Soares, E. A., Angelov, P. P., Costa, B., Castro, M., Nageshrao, S., & Filev, D. (2020). Explaining deep learning models through rule-based approximation and visualization. IEEE Transactions on Fuzzy Systems, 1, 1–10.

[40] Couteaux, V., Nempont, O., Pizaine, G., & Bloch, I. (2019). Towards interpretability of segmentation networks by analyzing DeepDreams. In Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support (pp. 56–63). Springer.

[41] Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4, eaao5580.

[42] Smith-Renner, A., Rua, R., & Colony, M. (2019). Towards an explainable threat detection tool. In IUI workshops.

[43] Stilgoe, J. (2020). Who Killed Elaine Herzberg? In: Who's Driving Innovation?. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-32320-2_1

[44] Soares, E., Angelov, P., Costa, B., & Castro, M. (2019). Actively semi-supervised deep rule-based classifier applied to adverse driving scenarios. In 2019 international joint conference on neural networks (IJCNN). IEEE (pp. 1–8).

[45] Jahmunah, V., Ng, E. Y. K., Tan, R. S., Oh, S. L., & Acharya, U. R. (2022). Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals. Computers in Biology and Medicine, 146, 105550. https://doi.org/10.1016/j.compbiomed.2022.105550

[46] Wei, K., Chen, B., Zhang, J., Fan, S., Wu, K., Liu, G., & Chen, D. (2022). Explainable Deep Learning Study for Leaf Disease Classification. Agronomy, 12, 1035. https://doi.org/10.3390/agronomy12051035

[47] De, T., Giri, P., Mevawala, A., Nemani, R., & Deo, A. (2020). Explainable AI: A Hybrid Approach to Generate Human-Interpretable Explanation for Deep

Learning Prediction. Procedia Computer Science, 168, 40-48. https://doi.org/10.1016/j.procs.2020.02.255

[48] Joshi, G., Walambe, R., & Kotecha, K. (2021). A Review on Explainability in Multimodal Deep Neural Nets. IEEE Access, 9, 59800-59821. doi: 10.1109/ACCESS.2021.3070212

[49] Naeem, H., Alshammari, B. M., & Ullah, F. (2022). Explainable Artificial Intelligence-Based IoT Device Malware Detection Mechanism Using Image Visualization and Fine-Tuned CNN-Based Transfer Learning Model. *Computational Intelligence and Neuroscience*, Volume 2022, Article ID 7671967, 17 pages. https://doi.org/10.1155/2022/7671967

[50] Roanec, J. M., Fortuna, B., Mladeni, D. (2022). Knowledge graph-based rich and confidentiality preserving Explainable Artificial Intelligence (XAI). Information Fusion, 81, 91-102. https://doi.org/10.1016/j.inffus.2021.11.015

[51] Kim, D., & Lee, J. (2022). Predictive evaluation of spectrogram-based vehicle sound quality via data augmentation and explainable artificial Intelligence: Image color adjustment with brightness and contrast. *Mechanical Systems and Signal Processing*, *Volume 179*, 109363. https://doi.org/10.1016/j.ymssp.2022.109363.

[52] Kakogeorgiou, I., & Karantzalos, K. (2021). Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, *Volume 103*, 102520. https://doi.org/10.1016/j.jag.2021.102520.

[53] Lim, S.-Y., Chae, D.-K., & Lee, S.-C. (2022). Detecting Deepfake Voice Using Explainable Deep Learning Techniques. *Appl. Sci.*, *12*, 3926. https://doi.org/10.3390/app12083926.

[54] Szczepański, M., Pawlicki, M., Kozik, R., et al. (2021). New explainability method for BERT-based model in fake news detection. *Sci Rep*, *11*, 23705. https://doi.org/10.1038/s41598-021-03100-6.

[55] Kim, H.-S., & Joe, I. (2022). An XAI method for convolutional neural networks in self-driving cars. *PLoS ONE*, *17*(8), e0267282. https://doi.org/10.1371/journal.pone.0267282.

[56] Reiter, E. (2019). Natural Language Generation Challenges for Explainable AI. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 3–7. Association for Computational Linguistics.

[57] Sai, A. M. A., Sai Eswar, K. L., Sai Harshith, K. S., Raghavendra, P., Kiran, G. Y., & M. V. (2022). Study of Lasso and Ridge Regression using ADMM. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, Hubli, India, pp. 1-8. doi: 10.1109/CONIT55038.2022.9847706.

[58] Van Deemter, K. (2010). *Not exactly: In praise of vagueness*. OUP Oxford.

[59] Daniel, K. (2017). *Thinking, fast and slow*.

[60] Reiter, E. (2019). Natural Language Generation Challenges for Explainable AI. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 3–7. Association for Computational Linguistics.

[61] Karakülah, G., Dicle, O., Koşaner, O., et al. (2014). Computer-based extraction of phenotypic features of human congenital anomalies from the digital literature with natural language processing techniques. *Stud Health Technol Inform*, *205*, 570–574.

[62] Dam, H. K., Tran, T., & Ghose, A. (2018). Explainable software analytics. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*, Gothenburg, Sweden, 27 May–3 June 2018, pp. 53–56.

[63] Lipton, Z. C. (2018). The mythos of model interpretability. *Commun. ACM*, *61*, 36–43.

[64] Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, *Volume 178*, *Issues 3–4*, 389–397. https://doi.org/10.1016/j.ecolmodel.2004.03.013.

[65] Adadi, A., & Berrada, M. (2020). Explainable AI for Healthcare: From Black Box to Interpretable Models. In *Advances in Intelligent Systems and Computing*, Volume 1076, pp. 327–337. Springer.

[66] Kleinbaum, D. G., & Kleinbaum, D. G. (1994). *Logistic Regression*. Springer.

[67] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.*, 2017, 4766–4775.

[68] Alvarez-Melis, D., & Jaakkola, T. S. (2017). A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 7–11 September 2017, pp. 412–421.