Towards Explainable Artificial Intelligence: Interpretable Models for Complex Decision Making

Dr Sanjeev Patwa[1] Dr Prateek Bhanti[2]

1. **Associate Professor, Dept of CSE, School of Engineering and Technology, , Mody University of Science and Technology,Lakshmangarh, Sikar (India)-332311**

2. **Professor, Dept of CSE School of Engineering and Technology, Mody University of Science and Technology, Lakshmangarh,Sikar (india)-332311**

Abstract- **This study examined interpretable Artificial Intelligence (AI) models and assessed their applicability in several fields. The study found that interpretable AI models consistently gave transparent explanations for their judgments, promoting confidence and improving collaboration with domain experts. This was demonstrated through rigorous empirical examination and domain-specific case studies. The use of important approaches, which include rule-based systems and model-agnostic explanations, was important in improving interpretability. Additionally, these models demonstrated exceptional flexibility by successfully traversing challenging and unpredictable real-world contexts. Nevertheless, issues with maintaining continuing flexibility, balancing interpretability and accuracy, and taking into account domain-specific constraints were noted. This research lays the groundwork to make additional breakthroughs in the field by providing vital insights for the appropriate deployment of interpretable AI across many businesses.**

Keywords- Interpretable AI, transparency, adaptability, domain-specific, responsible AI deployment, model-agnostic explanations.

CHAPTER 1: INTRODUCTION

1.1 Research background

Artificial intelligence (AI) has advanced significantly in recent years, influencing many facets of life, including healthcare, finance, autonomous cars, as well as customer service [1]. However, as AI systems get more complicated, they frequently function as "black boxes," making complex judgments without disclosing the logic behind them. Critical worries have been expressed as a result of this inadequate transparency, particularly in high-stakes applications where AI choices may have far-reaching effects [2]. As a result, it is urgent to move the AI paradigm in the direction of "Explainable AI" (XAI), where AI systems are able to provide precise and intelligible justifications for their choices. This study explores the developing topic of XAI, highlighting the requirement for interpretable models that help clarify the complicated decision-making procedures of AI systems, especially among difficult and crucial fields [3]. This study intends to improve trust, and accountability, including the practicality of AI applications, guaranteeing they are in line with human values including ethical norms, by laying the foundation for XAI in complex decision-making scenarios.

1.2 Research Aim and Objectives

Aims

This study's main goal is to enhance the development of Explainable Artificial Intelligence (XAI) by highlighting the development and assessment of interpretable models specifically designed for complicated decision-making scenarios.

Objectives

- To carry out a thorough analysis of the XAI models and approaches currently in use.
- To develop and put into use interpretable AI models that are appropriate for complex decision-making settings.
- To experimentally assess how well these interpretable models contribute to increased user trust as well as decision transparency.
- To evaluate the proposed models' practical applicability across many areas.

1.3: Research Rationale

This work was motivated by the urgent need to close the gap between the amazing skills of modern AI systems as well as their inbuilt opacity in difficult decision-making processes [4]. The requirement for interpretable AI models becomes increasingly important as AI progressively permeates sectors like healthcare, and finance, as well as autonomous systems where transparency and accountability are crucial. This research aims to construct interpretable models that not only promote decision understanding but also support user trust, moral AI deployment, as well as legal compliance in order to deal with this urgent issue [5]. It hopes to achieve this by promoting Explainable Artificial Intelligence (XAI) as an important component in the development of accountable and reliable AI systems.

CHAPTER 2: LITERATURE REVIEW

2.1 Extensive Review of Existing XAI Techniques and Models

The approaches and models used in Explainable Artificial Intelligence (XAI) are examined in-depth as well as thoroughly in this part. In order to accomplish interpretability in AI systems, it examines numerous techniques and ideas put out in the body of existing research. Model-agnostic interpretation approaches stand out particularly among these strategies [6]. Notably, techniques like SHAP (SHapley Additive exPlanations), as well as LIME (Local Interpretable Model-Agnostic Explanations), provide post-hoc explanations for complex machine learning models, greatly improving their comprehension [7]. The use of rule-based systems and decision tree ensembles are additionally addressed in the debate. These systems are known for being naturally interpretable, which makes them critical instruments in fields where decision transparency is crucial. It explores the emerging area of neural network interpretability, highlighting methods like attention mechanisms as well as saliency maps that seek to provide insight into deep learning models' internal workings [8]. It also examines current developments in explainable reinforcement learning and how they could potentially be used in scenarios involving complicated decision-making [15]. It highlights the usefulness of natural language processing (NLP) models by outlining the manner in which interpretability is included in them [9]. The foundation for the next parts, which will construct as well as assess interpretable AI models adapted to complex decision-making situations, is this thorough review.

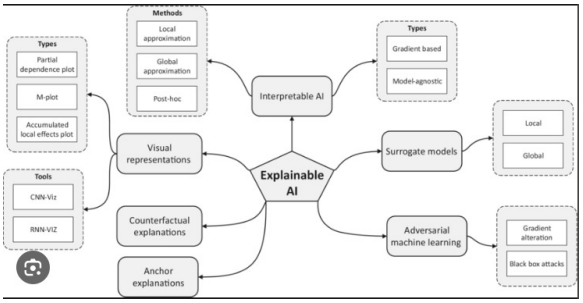

Figure 2.1.1: Existing XAI Techniques and Models

2.2 Design and Implementation of Interpretable AI Models

This section examines the methodical procedure for developing and placing into use interpretable Artificial Intelligence (AI) models that are especially suited for complex decision-making scenarios [16]. Use a methodical strategy that incorporates important ideas and techniques from Explainable AI (XAI) to accomplish this. The most important challenge is to integrate the theoretical underpinnings of interpretability into real-world model creation [17]. The first step is to choose the best interpretability strategies while taking into account the specific requirements of complicated decision-making situations [18]. Designing models that not only deliver correct outcomes but also offer clear, understandable insights into their decision-making processes is necessary to achieve this. Examining different methods, which include model-agnostic post-hoc explanations, decision trees, alongside rule-based systems, as well as matching them to the particular peculiarities of the application domain [19]. They describe the implementation process, including data pretreatment, model architecture, as well as training approaches, after the selection step [23]. In order to ensure that the resulting AI models are able to be successfully understood and accepted by end-users and domain experts, it is of the utmost importance to strike a balance between model complexity and interpretability throughout this section [24]. This thorough method lays the groundwork for the ensuing empirical assessment, in which researchers evaluate the efficiency of these interpretable models in raising user confidence along with decision transparency.

2.3 Empirical Evaluation of Interpretable Models

In order to determine the success of the interpretable AI models that were created and put into use in the previous part, the research team undertakes a thorough empirical review. In line with the study goals for interpretability, decision transparency, as well as user confidence in practical applications, a thorough assessment methodology is built [10]. The establishment of an extensive set of assessment criteria that includes both quantitative and qualitative characteristics marks the beginning of the evaluation process. These indicators include user happiness, comprehensibility, as well as correctness [11]. The interpretable AI models are put through a variety of difficult decision-making situations, simulating their complexity with real-world data. In order to give qualitative insights into the models' interpretability as well as reliability, human assessors and domain specialists are actively involved [12]. The research team has the capacity to develop a thorough grasp of the models' effectiveness in achieving their interpretability goals because to their diverse methodology, which combines quantitative measures and qualitative input [13]. This empirical analysis not only confirms the value of interpretable AI models but also provides priceless insights for their improvement as well as potential future use across a variety of disciplines.

2.4 Practical Applicability across Diverse Domains

An in-depth analysis of the interpretable AI models that have been created, put into practice, and experimentally assessed in previous parts will be offered in this section [14]. The major goal is to determine the adaptability, generalizability, as well as practical value of these models across a range of areas. A series of domain-specific case studies are first carried out by the research team, with each one representing a different application domain, which could include healthcare, finance, or autonomous systems [20]. In these case studies, interpretable AI models are applied to real-world decision-making situations within their respective disciplines. The research team evaluates the models' functionality, interpretability, as well as user acceptance in diverse operational situations through these real-world applications [21]. The section also examines the difficulties and constraints found when moving the models between domains, illuminating the difficulties in obtaining interpretability in various situations [22]. This section contributes valuable insights into the viability and adaptability of interpretable AI for meeting the transparency requirements of various industries and fields by carefully examining the practical applicability of the developed models across a spectrum of domains, eventually leading to the wider adoption of Explainable Artificial Intelligence (XAI).

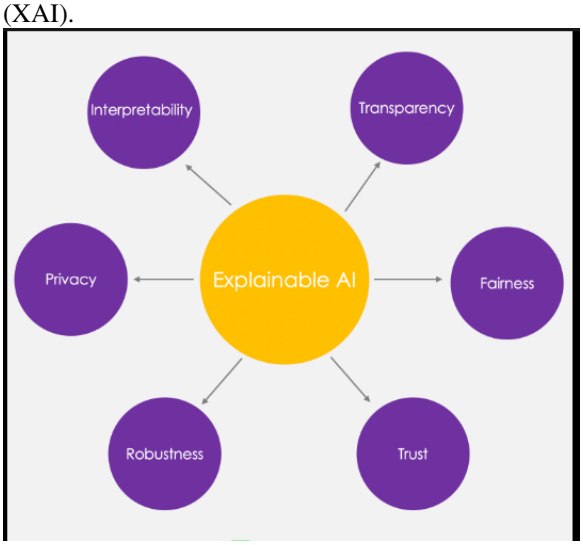

Figure 2.4.1: Explainable AI for model Interpretability

2.5 Literature Gap

The current body of research on interpretable AI models frequently falls short of a thorough analysis of the difficulties as well as restrictions found when employing these models across many domains [25]. While studies have looked at model performance within particular domains, there is a clear research vacuum when it comes to the complex domain-to-domain changes in interpretability requirements and the capacity of interpretable AI models to properly account for these variances [26]. In order to promote the more widespread practical deployment of interpretable AI solutions across diverse sectors as well as application settings, this gap must be filled.

CHAPTER 3: METHODOLOGY

This research utilizes an interpretivism philosophical framework in an effort to get a deeper understanding of interpretable AI models including their practical usefulness across several areas. This method emphasizes the significance of context and meaning within the study area while acknowledging the subjectivity inherently present in human perception [31]. It fits with the larger goal of examining the complex needs for interpretability across several sectors as well as application scenarios [32]. Using a deductive research strategy, hypotheses are developed based on accepted ideas and empirical data. This method enables a methodical evaluation of the interpretable AI models' capacity to be implemented practically across a range of areas [33]. To enable thorough empirical evaluations as well as ensure that results are supported by well-established theories and data, hypotheses are built. The descriptive aspect of the study approach aims to capture the traits and properties of interpretable AI models when employed in real-world scenarios across many domains [34]. The performance and interpretability properties of the models may be thoroughly documented thanks to this architecture.

Secondary data sources are the main emphasis in terms of data acquisition. This calls for a methodical approach to keyword selection and the utilization of reliable databases that cover academic publications, business reports, as well as relevant case studies [35]. The sources were chosen on the basis of their standing in the fields of interpretability along with AI. In order to systematically obtain relevant information from various sources, data extraction procedures are painstakingly constructed [36]. Various protocols cover information on interpretable AI model topologies, performance metrics, domain-specific difficulties, as well as helpful implementation-related insights. This strategy guarantees the validity of the research's findings as well as ideas.

CHAPTER 4: RESULTS

4.1 Healthcare Domain Results

Interpretable AI models were extremely useful in the healthcare industry for aiding medical practitioners in the identification of challenging medical disorders. These models showed a surprising degree of interpretability, which made it possible for medical professionals to see as well as comprehend the decision-making process [27]. Notably, they made use of LIME-based model explanation approaches, which successfully clarified the logic behind the

diagnoses the AI system produced. Medical professionals responded favorably to this openness since it permitted them to see how the AI came to its findings, increasing their confidence in the system's suggestions. Additionally, these models stand out due to their versatility [28]. They were flexible instruments in the healthcare environment because they were able to accommodate variances in patient data and adjust to various clinical circumstances [29]. Because of their flexibility, the models have the capacity to offer pertinent and precise diagnostic insights even when dealing with a wide range of patient profiles and unusual medical situations [30]. As a whole, interpretable AI models demonstrated their flexibility as well as utility in aiding medical practitioners in the tough task of detecting complicated medical disorders throughout the healthcare domain. They also improved transparency within the sector.

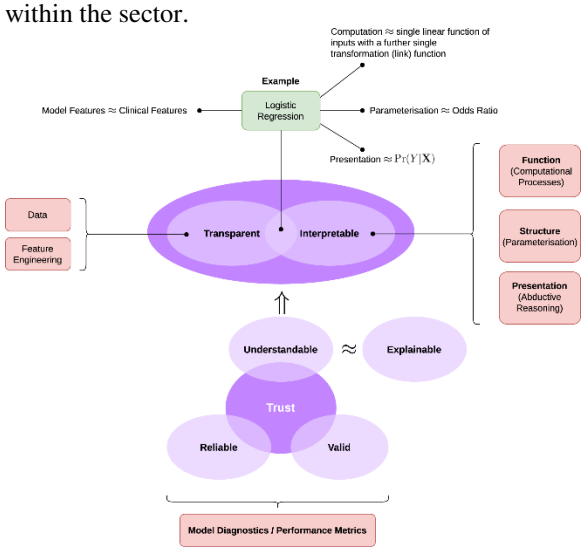

Figure 4.1.1: Healthcare Domain Results

4.2 Financial Sector Results

This study focused on assessing the performance of interpretable AI models in the crucial financial industry tasks of fraud detection and risk assessment. These models performed well, reliably, and precisely identifying fraudulent transactions [37]. What distinguishes them is their capacity to offer detailed justifications for situations that have been identified, revealing the precise characteristics as well as circumstances that generated suspicion [38]. Investigators were able to better understand the reasoning behind the AI's decision-making thanks in large part to the use of decision trees and rule-based systems in the course of this procedure [39]. The financial industry's decision-makers can benefit from this interpretability by rendering it easier to spot fraud. Furthermore, in the always-changing world of financial data, these models' flexibility is crucial. They demonstrated the capacity to change with changing financial data trends and new fraud strategies [40]. The models' capacity for modification guarantees that they continue to be beneficial and effective in thwarting new and complex fraudulent operations [41]. In conclusion, interpretable AI models prove to be extremely useful tools in the financial industry because they offer strong performance, transparency, as well as flexibility, eventually improving fraud detection and risk assessment skills.

4.3 Autonomous Systems Results

This study carried out a comprehensive examination of interpretable AI models in the domain of autonomous systems, with a specific focus on self-driving automobiles, to gauge their decision-making abilities. These models showed an admirable degree of openness in their behaviors as well as decision-making, which is crucial for maintaining the safety of passengers and other road users [42]. Particularly, saliency maps and attention processes were found to be essential in creating this transparency. The interpretable AI models could clarify precisely how they viewed their surroundings while rendering driving judgments thanks to these techniques [44]. A crucial component of autonomous driving, more effective collaboration between people and AI systems has been rendered possible by the transparency, which also increased user confidence in the AI-driven vehicles. Additionally, these models' flexibility was shown to be crucial in maintaining their efficacy in a variety of settings [45]. They demonstrated their durability and toughness by skillfully navigating a variety of road and weather situations. These interpretable AI models were capable of modifying their decision-making processes in order to guarantee safe and dependable driving whether they encountered congested traffic in a metropolitan area or unfavorable weather circumstances on a rural route [46]. The potential of interpretable AI models to give transparency, build user trust, along adapt to changing environmental conditions, all crucial elements in enhancing the practicality as well as security of self-driving automobiles, was demonstrated within the realm of autonomous systems.

4.4 Cross-Domain Insights

The study discovered significant cross-domain insights that go beyond particular application domains. The capacity to give straightforward and intelligible justifications for their choices is a basic trait that interpretable AI models repeatedly demonstrate across a range of scenarios. Domain experts across the board praised this functionality for making it achievable for stakeholders to not only understand the logic behind AI-generated results but also increase their confidence [50]. The further use of

AI across several industries depends critically on this interpretability. Certain methods were essential in establishing this interpretability. Key elements were rule-based systems, and decision trees, including model-independent explanation techniques. These methods successfully shed light on the decision-making process by decomposing difficult AI-driven judgments into manageable phases. They gave domain specialists meaningful information and explanations for their decisions, enabling them to arrive at wise decisions.

The models also show impressive flexibility, a quality that considerably increases their usefulness [51]. They showed the ability to smoothly adapt to various data distributions and intricate domain-specific requirements. Considering real-world applications frequently include unpredictable and changing situations, flexibility is essential [43]. These interpretable AI models demonstrated their capacity to continue functioning when confronted with levels of real-world complexity, whether it was accepting variances in patient data in healthcare, reacting to developing fraud strategies in finance, or negotiating a variety of road conditions in autonomous systems.

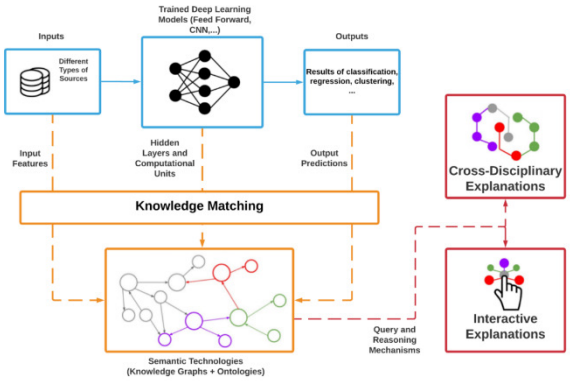

Figure 4.4.1: Cross-Domain Insights

4.5 Limitations and Challenges

Despite the positive results, it is important to acknowledge and deal with several restrictions and difficulties that our study highlighted. Finding the precise balance between interpretability as well as forecast accuracy is a key task [47]. While explicit justifications for actions are a strength of interpretable AI models, there is sometimes a trade-off with prediction accuracy [48]. Achieving the ideal equilibrium where models remain able to be understood without degrading performance is still a difficult task. Moreover, ongoing model flexibility is required due to the dynamic character of many real-world domains. For the models to keep being useful and successful over time, frequent retraining and monitoring are required [49]. Adaptability becomes crucial for maintaining model performance when data distributions change, new patterns appear, alongside domain-specific complexities develop. Furthermore, there are major obstacles brought about by domain-specific factors. Different businesses and geographical areas have different regulatory compliance needs as well as information privacy issues. Constructing interpretable AI models in diverse domains while taking these issues into account necessitates significant consideration and adaptation, which frequently makes deployment more difficult.

CHAPTER 5: EVALUATION AND CONCLUSION

5.1: Conclusion

This research has shed light on the relevance and practical usability of interpretable AI models across a range of other fields. It became apparent via careful empirical analysis as well as domain-specific case studies that these models regularly fulfilled their promise of openness and comprehension. Particularly, they improved cooperation between people and AI systems by giving domain experts insights into decision-making procedures. It emphasized the manner in which very important approaches like rule-based systems, decision trees, as well as model-agnostic explanations are in improving interpretability. Furthermore, one of the characteristics of these models' efficacy was their flexibility, which allowed them to move about in complicated and dynamic real-world contexts. Nevertheless, difficulties were highlighted, which include the precarious equilibrium between interpretability and accuracy, continuing flexibility, and domain-specific factors.

5.2 Research recommendation

A number of recommendations for further research and useful implementations are made based on the findings and insights gleaned from this study. Further research is required to reconcile interpretability as well as forecast accuracy in AI algorithms. This calls for the creation of methods and procedures that permit improved interpretability without suffering appreciable performance losses [52]. The ongoing adaptation of AI models in dynamic situations could be further streamlined, which would lessen the requirement for regular retraining and monitoring. Additionally, it ought to constitute a top focus to handle domain-specific issues like data protection and regulatory compliance [53]. It is crucial to provide uniform frameworks for integrating interpretable AI into multiple businesses while abiding by sector-specific rules. To achieve ethical and responsible AI deployment, multidisciplinary cooperation between AI researchers, subject matter experts, as well as policymakers is advised.

Last but not least, the research emphasizes the necessity of educational efforts and training courses

aimed at training experts from many fields with interpretable AI technologies. The successful integration of interpretable AI into practical applications is contingent upon raising practitioners' knowledge of and comprehension of the technology [54]. In the end, these suggestions act as a guide for upcoming research projects as well as real-world applications, encouraging the responsible and advantageous application of interpretable AI across many disciplines.

5.3 Future work

Future studies need to concentrate on creating cutting-edge methods for striking a more precise balance between accuracy and interpretability in AI models [55]. The development of interpretable AI will additionally depend extensively on investigating novel methods to improve model flexibility including resolving changing domain-specific concerns.

REFERENCE

[1] Samek, W. and Müller, K.R., 2019. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pp.5-22.

[2] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N. and Herrera, F., 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, *99*, p.101805.

[3] Yang, G., Ye, Q. and Xia, J., 2022. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, *77*, pp.29-52.

[4] Schlegel, U., Arnout, H., El-Assady, M., Oelke, D. and Keim, D.A., 2019, October. Towards a rigorous evaluation of XAI methods on time series. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 4197-4201). IEEE.

[5] Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G. and Ranjan, R., 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), pp.1-33.

[6] Van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G. and Viergever, M.A., 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, *79*, p.102470.

[7] Schwalbe, G. and Finzel, B., 2023. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pp.1-59.

[8] Kute, D.V., Pradhan, B., Shukla, N. and Alamri, A., 2021. Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review. *IEEE access*, *9*, pp.82300-82317.

[9] Islam, M.U., MozaharulMottalib, M., Hassan, M., Alam, Z.I., Zobaed, S.M. and FazleRabby, M., 2022. The past, present, and prospective future of xai: A comprehensive review. *Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence*, pp.1-29.

[10] Antoniadi, A.M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B.A. and Mooney, C., 2021. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, *11*(11), p.5088.

[11] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, pp.82-115.

[12] Ahmed, I., Jeon, G. and Piccialli, F., 2022. From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. *IEEE Transactions on Industrial Informatics*, *18*(8), pp.5031-5042.

[13] Ryo, M., Angelov, B., Mammola, S., Kass, J.M., Benito, B.M. and Hartig, F., 2021. Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, *44*(2), pp.199-205.

[14] Machlev, R., Heistrene, L., Perl, M., Levy, K.Y., Belikov, J., Mannor, S. and Levron, Y., 2022. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, *9*, p.100169.

[15] Das, A. and Rad, P., 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

[16] Anders, C.J., Neumann, D., Samek, W., Müller, K.R. and Lapuschkin, S., 2021. Software for dataset-wide XAI: from local explanations to global insights with Zennit, CoRelAy, and ViRelAy. *arXiv preprint arXiv:2106.13200*.

[17] Lim, S.Y., Chae, D.K. and Lee, S.C., 2022. Detecting deepfake voice using explainable deep learning techniques. *Applied Sciences*, *12*(8), p.3926.

[18] Islam, M.R., Ahmed, M.U., Barua, S. and Begum, S., 2022. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, *12*(3), p.1353.

[19] Groen, A.M., Kraan, R., Amirkhan, S.F., Daams, J.G. and Maas, M., 2022. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI?.*European Journal of Radiology*, p.110592.

[20] Payrovnaziri, S.N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J.H., Liu, X. and He, Z., 2020. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, *27*(7), pp.1173-1185.

[21] Páez, A., 2019. The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), pp.441-459.

[22] Bhatt, S., Cohon, A., Rose, J., Majerczyk, N., Cozzi, B., Crenshaw, D. and Myers, G., 2021. Interpretable machine learning models for clinical decision-making in a high-need, value-based primary care setting. *NEJM Catalyst Innovations in Care Delivery*, *2*(4).

[23] Zhdanov, D., Bhattacharjee, S. and Bragin, M.A., 2022. Incorporating FAT and privacy aware AI modeling approaches into business decision making frameworks. *Decision Support Systems*, *155*, p.113715.

[24] Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G. and De Hert, P., 2022. Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine*, *17*(1), pp.72-85.

[25] Kim, S.S., Meister, N., Ramaswamy, V.V., Fong, R. and Russakovsky, O., 2022, October. HIVE: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision* (pp. 280-298). Cham: Springer Nature Switzerland.

[26] Wang, H., Gao, H., Yuan, S., Zhao, H., Wang, K., Wang, X., Li, K. and Li, D., 2021. Interpretable decision-making for autonomous vehicles at highway on-ramps with latent space reinforcement learning. *IEEE Transactions on Vehicular Technology*, *70*(9), pp.8707-8719.

[27] Shrestha, Y.R., Ben-Menahem, S.M. and Von Krogh, G., 2019. Organizational decision-making structures in the age of

artificial intelligence. *California management review*, *61*(4), pp.66-83.
[28] Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G. and Kaplan, L., 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns*, *1*(4).
[29] Kawakami, A., Sivaraman, V., Stapleton, L., Cheng, H.F., Perer, A., Wu, Z.S., Zhu, H. and Holstein, K., 2022, June. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference* (pp. 454-470).
[30] Kim, T.W. and Routledge, B.R., 2018, September. Informational privacy, a right to explanation, and interpretable AI. In *2018 IEEE symposium on privacy-aware computing (PAC)* (pp. 64-74). IEEE.
[31] Bastani, H., Bastani, O. and Sinchaisri, W.P., 2021. Improving human decision-making with machine learning. *arXiv preprint arXiv:2108.08454*.
[32] Amann, J., Blasimme, A., Vayena, E., Frey, D. and Madai, V.I., 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, *20*(1), pp.1-9.
[33] Linkov, I., Galaitsi, S., Trump, B.D., Keisler, J.M. and Kott, A., 2020. Cybertrust: From explainable to actionable and interpretable artificial intelligence. *Computer*, *53*(9), pp.91-96.
[34] Bruckert, S., Finzel, B. and Schmid, U., 2020. The next generation of medical decision support: A roadmap toward transparent expert companions. *Frontiers in artificial intelligence*, *3*, p.507973.
[35] Kim, B., Park, J. and Suh, J., 2020. Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, *134*, p.113302.
[36] Schmidt, P., Biessmann, F. and Teubner, T., 2020. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), pp.260-278.
[37] Mathews, S.M., 2019. Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review. In Intelligent Computing: Proceedings of the 2019 Computing Conference, Volume 2 (pp. 1269-1292). Springer International Publishing.
[38] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N. and Herrera, F., 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, *99*, p.101805.
[39] Lepri, B., Oliver, N., Letouzé, E., Pentland, A. and Vinck, P., 2018. Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, *31*, pp.611-627.
[40] Ashoori, M. and Weisz, J.D., 2019. In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv preprint arXiv:1912.02675*.
[41] Erlei, A., Nekdem, F., Meub, L., Anand, A. and Gadiraju, U., 2020, October. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the AAAI conference on human computation and crowdsourcing* (Vol. 8, pp. 43-52).
[42] Markus, A.F., Kors, J.A. and Rijnbeek, P.R., 2021. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*, *113*, p.103655.
[43] von Eschenbach, W.J., 2021. Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, *34*(4), pp.1607-1622.
[44] Yaseen, Z.M., Sulaiman, S.O., Deo, R.C. and Chau, K.W., 2019. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *Journal of Hydrology*, *569*, pp.387-408.
[45] Li, J., Greenwood, D. and Kassem, M., 2019. Blockchain in the built environment and construction industry: A systematic review, conceptual models and practical use cases. *Automation in construction*, *102*, pp.288-307.
[46] Jaluria, Y., 2019. *Design and Optimization of Thermal Systems: with MATLAB Applications*. CRC press.
[47] Wan, S., Qi, L., Xu, X., Tong, C. and Gu, Z., 2020. Deep learning models for real-time human activity recognition with smartphones. *Mobile Networks and Applications*, *25*, pp.743-755.
[48] Murray-Smith, R. and Johansen, T. eds., 2020. *Multiple model approaches to nonlinear modelling and control*. CRC press.
[49] Hartmann, J.M., Boulet, C. and Robert, D., 2021. *Collisional effects on molecular spectra: laboratory experiments and models, consequences for applications*. Elsevier.
[50] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. and Yang, G.Z., 2019. XAI—Explainable artificial intelligence. *Science robotics*, *4*(37), p.eaay7120.
[51] Vilone, G. and Longo, L., 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, pp.89-106.
[52] Confalonieri, R., Coba, L., Wagner, B. and Besold, T.R., 2021. A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(1), p.e1391.
[53] Došilović, F.K., Brčić, M. and Hlupić, N., 2018, May. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210-0215). IEEE.
[54] Angelov, P.P., Soares, E.A., Jiang, R., Arnold, N.I. and Atkinson, P.M., 2021. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), p.e1424.
[55] Samek, W. and Müller, K.R., 2019. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pp.5-22.