Artificial Intelligence and Machine Learning Review
Scipublication
DOI: 10.69987/AIMLR.2021.20101

Explainable AI (XAI): Methods, Tools, and Challenges in Interpreting Machine Learning Models

Ravi Shankar[1], *Farhana Ahmed*[2]
Department of Computer Science, University of Rajshahi[1], *Bangladesh, Faculty of Engineering, Khulna University, Bangladesh*[2]
ravi.shankar@ru.ac.bd[1], farhana.ahmed@ku.ac.bd[2]

Keywords

Explainable AI, Machine Learning, Interpretability, Transparency, Model Explainability, XAI Tools, Ethical AI, Model Complexity

Abstract

Explainable AI (XAI) has emerged as a pivotal area of research in artificial intelligence (AI) and machine learning (ML), addressing the growing need for transparency, interpretability, and accountability in AI systems. As machine learning models become increasingly complex and pervasive, their "black-box" nature poses significant challenges, particularly in high-stakes domains such as healthcare, finance, and criminal justice. This research article provides a comprehensive exploration of XAI, focusing on the methods, tools, and challenges associated with interpreting machine learning models. The article begins by discussing the importance of explainability in AI, emphasizing its role in building trust, ensuring accountability, and enabling human oversight. It then delves into various techniques for achieving explainability, including model-specific methods (e.g., decision trees, rule-based models) and model-agnostic approaches (e.g., LIME, SHAP). The article also highlights the tools available for implementing these techniques, ranging from open-source libraries like LIME and SHAP to commercial platforms such as IBM Watson Open Scale and Google Cloud Explainable AI. Furthermore, the article addresses the challenges and limitations of XAI, including ethical considerations, trade-offs between accuracy and interpretability, and the lack of standardized evaluation metrics. The article concludes with a discussion of future directions for XAI research, emphasizing its potential to transform industries by making AI systems more transparent, interpretable, and trustworthy. This work aims to serve as a foundational resource for researchers and practitioners seeking to advance the field of explainable AI.

1. Introduction

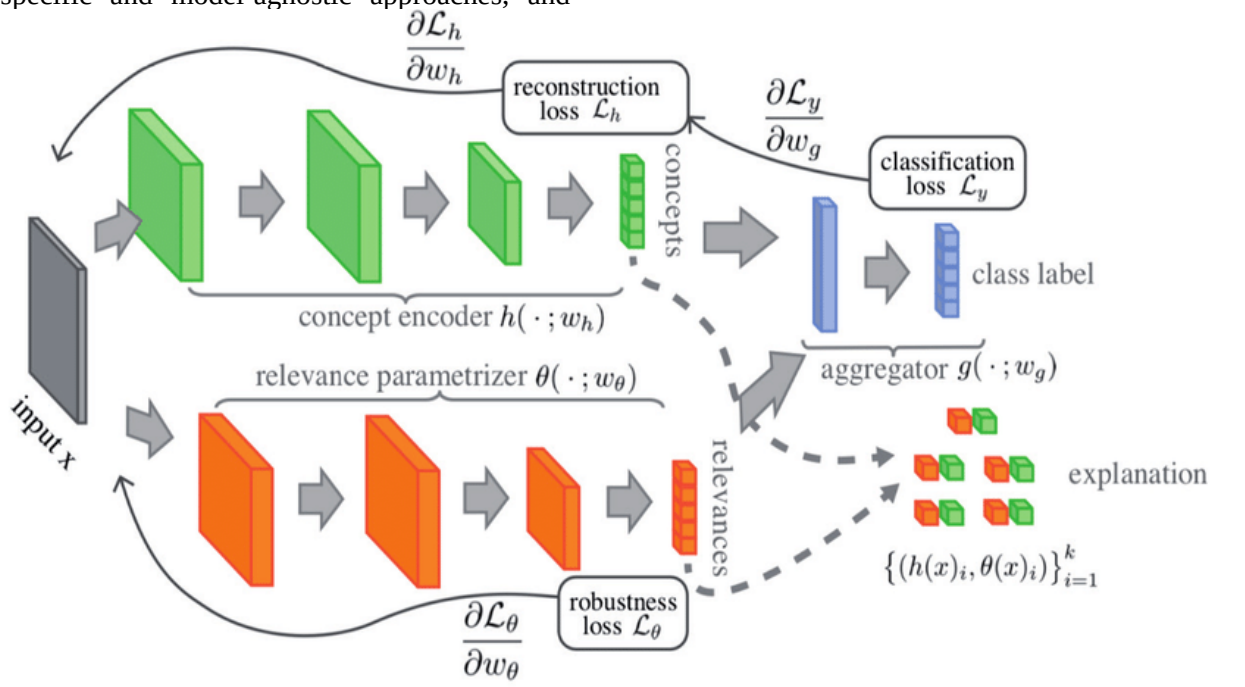
1.1 Background and Motivation

The rapid advancement of machine learning (ML) and artificial intelligence (AI) has led to the development of highly complex models that can achieve state-of-the-art performance across a wide range of tasks. However, as these models become more sophisticated, they also become increasingly opaque, making it difficult for humans to understand how they arrive at their predictions. This lack of transparency poses significant challenges, particularly in high-stakes domains such as healthcare, finance, and criminal justice, where the consequences of incorrect or biased decisions can be severe.

Explainable AI (XAI) aims to address these challenges by developing methods and tools that make the decision-making processes of AI systems more transparent and interpretable[1]. The goal of XAI is not only to improve the trustworthiness of AI systems but also to enable humans to understand, validate, and ultimately control these systems. This is particularly important as AI systems are increasingly being used to make critical decisions that affect individuals and society as a whole[2].

1.2 Objectives and Scope

The primary objective of this research article is to provide a comprehensive overview of the current state of XAI, with a focus on the methods, tools, and challenges associated with interpreting machine learning models. The article will explore various techniques for achieving explainability, including model-specific and model-agnostic approaches, and discuss the tools available for implementing these techniques. Additionally, the article will highlight the challenges and limitations of XAI, including ethical considerations, trade-offs between accuracy and interpretability, and the need for standardized evaluation metrics[3].



The scope of this article is broad, encompassing both theoretical and practical aspects of XAI. The article will cover a wide range of topics, including the importance of explainability in AI, the different types of explanations that can be generated, and the various methods and tools available for achieving explainability. The article will also discuss the challenges and limitations of XAI, as well as future directions for research in this area[4].

1.3 Structure of the Article

The remainder of this article is organized as follows. Section 2 provides an overview of the importance of explainability in AI and the different types of explanations that can be generated. Section 3 explores various methods for achieving explainability, including model-specific and model-agnostic approaches. Section 4 discusses the tools available for implementing these methods, with a focus on open-source libraries and platforms. Section 5 highlights the challenges and

limitations of XAI, including ethical considerations, trade-offs between accuracy and interpretability, and the need for standardized evaluation metrics. Section 6 concludes the article with a discussion of future

directions for XAI research and its potential impact on various industries[5].

2. The Importance of Explainability in AI

2.1 The Need for Transparency and Interpretability

As AI systems become more pervasive, the need for transparency and interpretability in these systems has grown exponentially. In many domains, the decisions made by AI systems can have significant consequences for individuals and society as a whole. For example, in healthcare, AI systems are being used to diagnose diseases, recommend treatments, and predict patient outcomes. In finance, AI systems are being used to assess creditworthiness, detect fraud, and make investment decisions. In criminal justice, AI systems are being used to predict recidivism, assess risk, and inform sentencing decisions[6].

Table 1: Comparison of Model-Specific and Model-Agnostic Methods for Explainability

Method Type	Examples	Strengths	Limitations
Model-Specific	Decision Trees, Rule-Based Models, Linear Models	Inherently interpretable, easy to visualize	Limited to specific model architectures
Model-Agnostic	LIME, SHAP, Partial Dependence Plots	Applicable to any model, flexible	May require additional computational resources
Hybrid	Anchors, Integrated Gradients, Counterfactual Explanations	Combines strengths of both approaches	May be complex to implement

In these and other high-stakes domains, it is essential that the decisions made by AI systems are transparent and interpretable. Without transparency, it is difficult for humans to understand how these systems arrive at their predictions, which can lead to mistrust and skepticism. Without interpretability, it is difficult for humans to validate the decisions made by these systems, which can lead to incorrect or biased outcomes[7].

2.2 Types of Explanations in AI

There are several types of explanations that can be generated by AI systems, each of which serves a different purpose. The most common types of explanations include:

Global Explanations: Global explanations provide an overview of how a model makes decisions across the entire dataset. These explanations are useful for understanding the overall behavior of a model and identifying any biases or patterns in the data[8].

Local Explanations: Local explanations provide insights into how a model makes decisions for individual instances or predictions. These explanations are useful for understanding why a model made a specific prediction and for identifying any anomalies or outliers in the data.

Model-Specific Explanations: Model-specific explanations are tailored to the specific architecture and parameters of a particular model. These explanations are useful for understanding the inner workings of a model and for identifying any specific features or parameters that are driving the model's predictions.

Model-Agnostic Explanations: Model-agnostic explanations are not tied to any specific model architecture or parameters. These explanations are useful for comparing the behavior of different models and for understanding the general principles that underlie a model's predictions[9].

2.3 The Role of Explainability in Trust and Accountability

Explainability plays a critical role in building trust and accountability in AI systems. When AI systems are transparent and interpretable, humans are more likely to trust the decisions made by these systems. This is particularly important in high-stakes domains, where the consequences of incorrect or biased decisions can be severe[10].

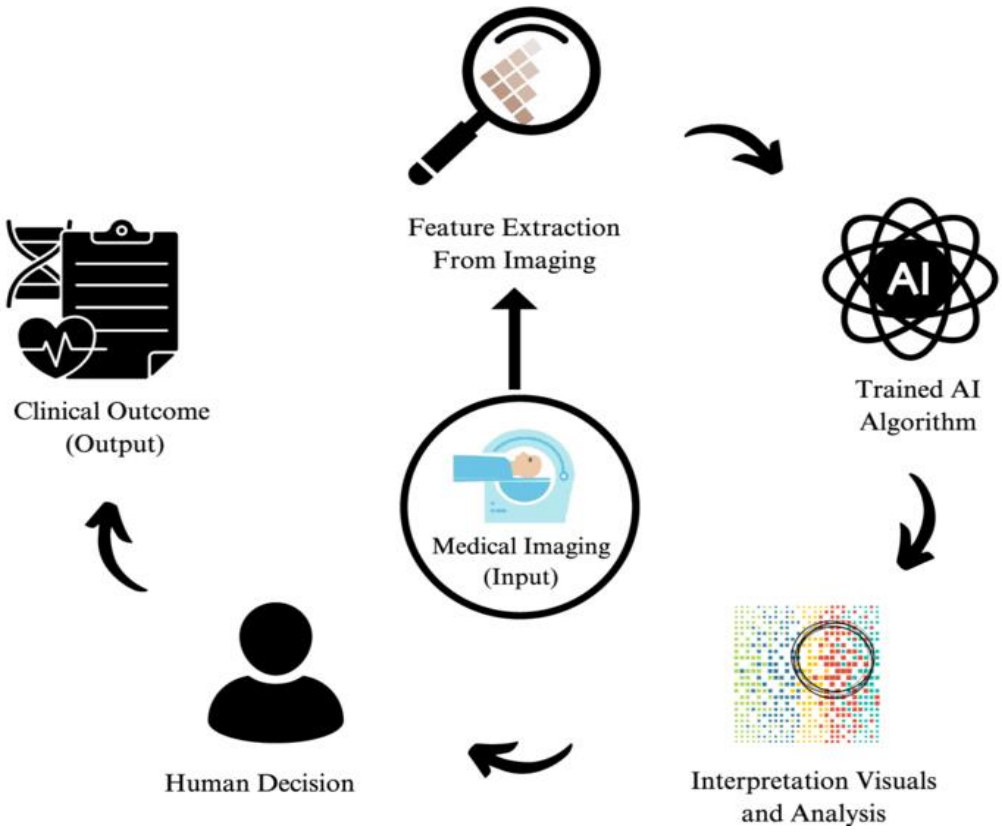
Explainability also enables humans to validate the decisions made by AI systems, which is essential for ensuring accountability. When humans can understand how a model arrives at its predictions, they can identify any errors or biases in the model and take corrective action. This is particularly important in domains where the decisions made by AI systems can have significant consequences for individuals and society as a whole[11].

3. Methods for Achieving Explainability in AI

3.1 Model-Specific Methods

Model-specific methods for achieving explainability are tailored to the specific architecture and parameters of a particular model. These methods are useful for understanding the inner workings of a model and for identifying any specific features or parameters that are driving the model's predictions. Some of the most common model-specific methods include:

Decision Trees: Decision trees are a type of model that is inherently interpretable. Each node in the tree represents a decision based on a specific feature, and the branches represent the possible outcomes of that decision. Decision trees are easy to visualize and understand, making them a popular choice for applications where interpretability is important[12].



Rule-Based Models: Rule-based models are another type of model that is inherently interpretable. These models use a set of predefined rules to make decisions, and the rules can be easily understood and validated by humans. Rule-based models are often used in domains where transparency and interpretability are critical, such as healthcare and finance[13].

Linear Models: Linear models are a type of model that is relatively simple and interpretable. These models make predictions based on a linear combination of input features, and the coefficients of the model can be easily interpreted as the importance of each feature. Linear models are often used in domains where interpretability is important, such as economics and social sciences.

3.2 Model-Agnostic Methods

Model-agnostic methods for achieving explainability are not tied to any specific model architecture or parameters. These methods are useful for comparing the behavior of different models and for understanding the general principles that underlie a model's predictions. Some of the most common model-agnostic methods include:

LIME (Local Interpretable Model-agnostic Explanations): LIME is a popular method for generating local explanations for individual predictions. The method works by approximating the behavior of a complex model with a simpler, interpretable model in the vicinity of a specific prediction. LIME is useful for understanding why a model made a specific prediction and for identifying any anomalies or outliers in the data.

SHAP (SHapley Additive exPlanations): SHAP is a method for generating both local and global explanations based on Shapley values from cooperative game theory. The method assigns a value to each feature that represents its contribution to the prediction, and these values can be aggregated to provide a global explanation of the model's behavior. SHAP is useful for understanding the overall behavior of a model and for identifying any biases or patterns in the data[14].

Partial Dependence Plots (PDPs): PDPs are a method for visualizing the relationship between a specific feature and the predicted outcome, while marginalizing over the effects of other features. PDPs are useful for understanding the impact of individual features on the model's predictions and for identifying any interactions between features[15].

3.3 Hybrid Methods

Hybrid methods for achieving explainability combine model-specific and model-agnostic approaches to provide a more comprehensive understanding of a model's behavior[16]. These methods are useful for leveraging the strengths of both approaches and for addressing the limitations of each. Some of the most common hybrid methods include:

Anchors: Anchors is a method for generating high-precision rule-based explanations for individual predictions. The method works by identifying a set of conditions (or "anchors") that are sufficient to guarantee a specific prediction with high confidence. Anchors is useful for understanding why a model made a specific prediction and for identifying any conditions that are critical to the model's decision-making process.

Integrated Gradients: Integrated Gradients is a method for attributing the prediction of a model to its input features. The method works by integrating the gradients of the model's output with respect to its input features along a path from a baseline input to the actual input. Integrated Gradients is useful for understanding the contribution of individual features to the model's predictions and for identifying any features that are driving the model's behavior[17].

Counterfactual Explanations: Counterfactual explanations are a method for generating "what-if" scenarios that show how the model's predictions would change if certain features were altered. These explanations are useful for understanding the sensitivity of the model's predictions to changes in the input features and for identifying any features that are critical to the model's decision-making process[18].

4. Tools for Implementing Explainability in AI

4.1 Open-Source Libraries and Platforms

There are several open-source libraries and platforms available for implementing explainability in AI. These tools provide a wide range of functionalities for generating explanations, visualizing model behavior, and evaluating the interpretability of models. Some of the most popular open-source libraries and platforms include:

LIME: LIME is an open-source Python library for generating local explanations for individual predictions. The library provides a simple and intuitive interface for approximating the behavior of complex models with simpler, interpretable models. LIME is widely used in both research and industry for understanding the behavior of black-box models[19].

SHAP: SHAP is an open-source Python library for generating both local and global explanations based on Shapley values. The library provides a wide range of functionalities for visualizing and interpreting the contributions of individual features to the model's predictions. SHAP is widely used in both research and industry for understanding the overall behavior of models and for identifying any biases or patterns in the data[20].

ELI5: ELI5 is an open-source Python library for explaining the predictions of machine learning models. The library provides a wide range of functionalities for generating explanations, including feature importance, partial dependence plots, and decision tree visualization. ELI5 is widely used in both research and industry for understanding the behavior of models and for debugging and improving model performance.

InterpretML: InterpretML is an open-source Python library for building interpretable models and explaining black-box models. The library provides a wide range of functionalities for generating explanations, including rule-based models, decision trees, and linear models. InterpretML is widely used in both research and industry for building transparent and interpretable models[21].

4.2 Commercial Tools and Platforms

In addition to open-source libraries and platforms, there are several commercial tools and platforms available for implementing explainability in AI. These tools provide a wide range of functionalities for generating explanations, visualizing model behavior, and evaluating the interpretability of models. Some of the most popular commercial tools and platforms include:

IBM Watson OpenScale: IBM Watson OpenScale is a commercial platform for monitoring, explaining, and managing AI models. The platform provides a wide range of functionalities for generating explanations, including feature importance, partial dependence plots, and counterfactual explanations. IBM Watson OpenScale is widely used in industry for ensuring the transparency and accountability of AI models[22].

Google Cloud Explainable AI: Google Cloud Explainable AI is a commercial platform for explaining the predictions of machine learning models. The platform provides a wide range of functionalities for generating explanations, including feature importance, Shapley values, and integrated gradients. Google Cloud Explainable AI is widely used in industry for understanding the behavior of models and for ensuring the transparency and accountability of AI models[23].

Microsoft Azure InterpretML: Microsoft Azure InterpretML is a commercial platform for building interpretable models and explaining black-box models. The platform provides a wide range of functionalities for generating explanations, including rule-based

models, decision trees, and linear models. Microsoft Azure InterpretML is widely used in industry for building transparent and interpretable models.

4.3 Case Studies and Applications

There are several case studies and applications of explainability in AI across various industries. These case studies demonstrate the importance of explainability in ensuring the transparency, interpretability, and accountability of AI systems. Some of the most notable case studies and applications include:

Healthcare: In healthcare, explainability is critical for ensuring the transparency and accountability of AI systems used for diagnosing diseases, recommending treatments, and predicting patient outcomes. For example, the LIME and SHAP libraries have been used to explain the predictions of machine learning models used for diagnosing breast cancer and predicting patient outcomes in intensive care units.

Finance: In finance, explainability is critical for ensuring the transparency and accountability of AI systems used for assessing creditworthiness, detecting fraud, and making investment decisions. For example, the ELI5 and InterpretML libraries have been used to explain the predictions of machine learning models used for assessing credit risk and detecting fraudulent transactions[24].

Table 2: Popular Tools and Platforms for Implementing Explainability in AI

Tool/Platform	Type	Key Features	Use Cases
LIME	Open-Source	Local explanations, model-agnostic	Understanding individual predictions
SHAP	Open-Source	Local and global explanations, Shapley values	Understanding overall model behavior
ELI5	Open-Source	Feature importance, partial dependence plots	Debugging and improving model performance
InterpretML	Open-Source	Interpretable models, rule-based models	Building transparent models
IBM Watson OpenScale	Commercial	Feature importance, counterfactual explanations	Ensuring transparency and accountability
Google Cloud Explainable AI	Commercial	Feature importance, Shapley values, integrated gradients	Understanding model behavior
Microsoft Azure InterpretML	Commercial	Interpretable models, rule-based models	Building transparent models

Criminal Justice: In criminal justice, explainability is critical for ensuring the transparency and accountability of AI systems used for predicting recidivism, assessing risk, and informing sentencing decisions. For example, the SHAP and InterpretML libraries have been used to explain the predictions of machine learning models used for predicting recidivism and assessing the risk of reoffending[25].

5. Challenges and Limitations of Explainable AI

5.1 Ethical Considerations

One of the most significant challenges in the field of explainable AI is the ethical considerations associated with the use of AI systems. As AI systems become more pervasive, the potential for these systems to be used in ways that are harmful or discriminatory increases. For example, AI systems used in criminal justice have been criticized for perpetuating racial biases, while AI systems used in hiring have been criticized for perpetuating gender biases[26].

Explainability plays a critical role in addressing these ethical considerations by enabling humans to understand and validate the decisions made by AI systems. However, achieving explainability is not always straightforward, particularly in cases where the models are highly complex or the data is highly sensitive. In these cases, it may be necessary to develop new methods and tools for achieving explainability that take into account the ethical implications of the decisions made by AI systems[27].

5.2 Trade-offs Between Accuracy and Interpretability

Another significant challenge in the field of explainable AI is the trade-offs between accuracy and interpretability. In many cases, the most accurate models are also the most complex and opaque, making it difficult to achieve explainability without sacrificing accuracy. For example, deep learning models, which are

known for their high accuracy, are often criticized for their lack of interpretability[28].

Achieving a balance between accuracy and interpretability is a key challenge in the field of explainable AI. In some cases, it may be necessary to sacrifice some degree of accuracy in order to achieve a higher level of interpretability. In other cases, it may be possible to develop new methods and tools that achieve both high accuracy and high interpretability. However, achieving this balance is not always straightforward, and it often requires a deep understanding of both the model and the data.

5.3 The Need for Standardized Evaluation Metrics

A third significant challenge in the field of explainable AI is the need for standardized evaluation metrics. Currently, there is no consensus on how to evaluate the explainability of AI systems, making it difficult to compare different methods and tools. This lack of standardized evaluation metrics is a significant barrier to progress in the field, as it makes it difficult to determine which methods and tools are most effective.

Developing standardized evaluation metrics for explainable AI is a key challenge that needs to be addressed in order to advance the field. These metrics should take into account both the quality of the explanations generated by the AI system and the impact of these explanations on the decision-making process. Additionally, these metrics should be applicable across a wide range of domains and use cases, making it possible to compare the explainability of different AI systems in a meaningful way.

6. Conclusion and Future Directions

6.1 Summary of Key Findings

This research article has provided a comprehensive overview of the current state of explainable AI (XAI), with a focus on the methods, tools, and challenges associated with interpreting machine learning models. The article has explored various techniques for achieving explainability, including model-specific and model-agnostic approaches, and discussed the tools available for implementing these techniques. Additionally, the article has highlighted the challenges and limitations of XAI, including ethical considerations, trade-offs between accuracy and interpretability, and the need for standardized evaluation metrics[29].

6.2 Future Directions for XAI Research

The field of explainable AI is still in its early stages, and there are many open questions and challenges that need to be addressed in order to advance the field. Some of the most promising directions for future research include:

Developing New Methods for Achieving Explainability: There is a need for new methods and tools that can achieve both high accuracy and high interpretability, particularly in cases where the models are highly complex or the data is highly sensitive. These methods should take into account the ethical implications of the decisions made by AI systems and should be applicable across a wide range of domains and use cases.

Standardizing Evaluation Metrics for Explainability: There is a need for standardized evaluation metrics that can be used to compare the explainability of different AI systems in a meaningful way. These metrics should take into account both the quality of the explanations generated by the AI system and the impact of these explanations on the decision-making process[30].

Exploring the Ethical Implications of XAI: There is a need for further research into the ethical implications of explainable AI, particularly in cases where the decisions made by AI systems can have significant consequences for individuals and society as a whole. This research should focus on developing methods and tools that can ensure the transparency, interpretability, and accountability of AI systems, while also taking into account the potential for these systems to be used in ways that are harmful or discriminatory.

6.3 The Potential Impact of XAI on Various Industries

Explainable AI has the potential to have a significant impact on various industries, including healthcare, finance, and criminal justice. By making the decision-making processes of AI systems more transparent and interpretable, XAI can improve the trustworthiness of these systems and enable humans to understand, validate, and ultimately control these systems. This is particularly important in high-stakes domains, where the consequences of incorrect or biased decisions can be severe.

In healthcare, XAI can improve the transparency and accountability of AI systems used for diagnosing diseases, recommending treatments, and predicting patient outcomes. In finance, XAI can improve the transparency and accountability of AI systems used for assessing creditworthiness, detecting fraud, and making investment decisions. In criminal justice, XAI can improve the transparency and accountability of AI systems used for predicting recidivism, assessing risk, and informing sentencing decisions.

Overall, the potential impact of XAI on various industries is significant, and the development of new methods and tools for achieving explainability is critical for realizing this potential. As the field of explainable AI continues to evolve, it is likely that we will see even

greater advancements in the transparency, interpretability, and accountability of AI systems, leading to improved outcomes for individuals and society as a whole[31].

References

[1] V. Nagisetty, L. Graves, J. Scott, and V. Ganesh, "XAI-GAN: Enhancing generative Adversarial Networks via explainable AI systems," *arXiv [cs.LG]*, 24-Feb-2020.

[2] E. Oduor, K. Qian, Y. Li, and L. Popa, "XAIT," in *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, Cagliari Italy, 2020.

[3] A. Ortega, J. Fierrez, A. Morales, Z. Wang, and T. Ribeiro, "Symbolic AI for XAI: Evaluating LFIT inductive programming for fair and explainable automatic recruitment," *arXiv [cs.AI]*, 01-Dec-2020.

[4] D. R. Chittajallu *et al.*, "XAI-CBIR: Explainable AI system for content based retrieval of video frames from minimally invasive surgery videos," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Venice, Italy, 2019.

[5] E. Molina-Perez, O. A. Esquivel-Flores, and H. Zamora-Maldonado, "Computational intelligence for studying sustainability challenges: Tools and methods for dealing with deep uncertainty and complexity," *Front. Robot. AI*, vol. 7, p. 111, Sep. 2020.

[6] J. Fu, C. Pan, and W. Xiong, "Intelligent scheduling methods for challenges of cluster tools with concurrent processing of multiple wafer types," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, Hong Kong, China, 2020.

[7] E. Thomann and M. Maggetti, "Designing research with qualitative comparative analysis (QCA): Approaches, challenges, and tools," *Sociol. Methods Res.*, vol. 49, no. 2, pp. 356–386, May 2020.

[8] M. Norqulova, "Ingliz tili grammatikasini akt bilan o"qitish," *Ренессанс в парадигме новаций образования и технологий в XXI веке*, no. 1, pp. 179–180, May 2022.

[9] Vaidehi and V. Vinod, "A survey on big data & deep learning: Methods, tools, applications, challenges and future trend," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 0009-SPECIAL, pp. 988–992, Sep. 2019.

[10] D. Dauletova and G. Dauletbaeva, "Self-study in learning foreign language," *Ренессанс в парадигме новаций образования и технологий в XXI веке*, no. 1, pp. 191–193, May 2022.

[11] M. Natonek-Wiśniewska and P. Krzyścin, "Detection of the species composition of food using mitochondrial DNA: Challenges and possibilities of a modern laboratory," in *Biochemical Analysis Tools - Methods for Bio-Molecules Studies*, IntechOpen, 2020.

[12] S. Saloustros, M. Cervera, and L. Pelà, "Challenges, tools and applications of tracking algorithms in the numerical modelling of cracks in concrete and masonry structures," *Arch. Comput. Methods Eng.*, vol. 26, no. 4, pp. 961–1005, Sep. 2019.

[13] A. Sadovykh *et al.*, "MegaM@Rt2 project: Mega-modelling at runtime - intermediate results and research challenges," in *Software Technology: Methods and Tools*, Cham: Springer International Publishing, 2019, pp. 393–405.

[14] A. Sadovykh *et al.*, "REVaMP2 project: Towards round-trip engineering of software product lines - approach, intermediate results and challenges," in *Software Technology: Methods and Tools*, Cham: Springer International Publishing, 2019, pp. 406–417.

[15] W. Bogaerts and L. Chrostowski, "Silicon photonics circuit design: Methods, tools and challenges," *Laser Photon. Rev.*, vol. 12, no. 4, p. 1700237, Apr. 2018.

[16] C. R. de Sá, W. Duivesteijn, P. Azevedo, A. M. Jorge, C. Soares, and A. Knobbe, "Discovering a taste for the unusual: exceptional models for preference mining," *Mach. Learn.*, vol. 107, no. 11, pp. 1775–1807, Nov. 2018.

[17] A. Bekesi *et al.*, "Challenges in the structural-functional characterization of multidomain, partially disordered proteins CBP and p300: Preparing native proteins and developing nanobody tools," *Methods Enzymol.*, vol. 611, pp. 607–675, Nov. 2018.

[18] P. Gamallo, "Using the outlier detection task to evaluate distributional semantic models," *Mach. Learn. Knowl. Extr.*, vol. 1, no. 1, pp. 211–223, Nov. 2018.

[19] M. Klaus and C. Genzel, "Residual stresses in thin films and coated tools: Challenges and strategies for their nondestructive analysis by X-ray diffraction methods," in *Neutrons and Synchrotron Radiation in Engineering Materials Science*, Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2017, pp. 439–449.

[20] S. Joel, P. W. Eastwick, and E. J. Finkel, "Open sharing of data on close relationships and other sensitive social psychological topics: Challenges, tools, and future directions," *Adv. Methods Pract. Psychol. Sci.*, vol. 1, no. 1, pp. 86–94, Mar. 2018.

[21] D. Bhatt *et al.*, "Opportunities and challenges for formal methods tools in the certification of avionics software," in *2017 IEEE Aerospace Conference*, Big Sky, MT, 2017.

[22] M. Garofalo, A. Botta, and G. Ventre, "Astrophysics and Big Data: Challenges, methods, and tools," *arXiv [astro-ph.IM]*, 15-Mar-2017.

[23] S. Dobbie, J. G. Dyke, and K. Schreckenberg, “The use of participatory methods & simulation tools to understand the complexity of rural food security,” in *Sustainable Development Challenges in the Arab States of the Gulf*, Gerlach Press, 2017, pp. 170–192.

[24] B. P. Evans, B. Xue, and M. Zhang, “What’s inside the black-box?,” in *Proceedings of the Genetic and Evolutionary Computation Conference*, Prague Czech Republic, 2019.

[25] A. Chatzimparmpas, R. M. Martins, I. Jusufi, and A. Kerren, “A survey of surveys on the use of visualization for interpreting machine learning models,” *Inf. Vis.*, vol. 19, no. 3, pp. 207–233, Jul. 2020.

[26] M. Ruffini, M. Casanellas, and R. Gavaldà, “A new method of moments for latent variable models,” *Mach. Learn.*, vol. 107, no. 8–10, pp. 1431–1455, Sep. 2018.

[27] M. Jaiswal, “Interpreting multimodal machine learning models trained for emotion recognition to address robustness and privacy concerns,” *Proc. Conf. AAAI Artif. Intell.*, vol. 34, no. 10, pp. 13716–13717, Apr. 2020.

[28] K. Weterings, S.-L. Kimelman, S. Bromuri, and M. van Eekelen, “Interpreting Attention Models: LSTM vs. CNN : A case study on customer activation,” in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Miami, FL, USA, 2020.

[29] E. Sood, S. Tannert, D. Frassinelli, A. Bulling, and N. T. Vu, “Interpreting attention models with human visual attention in machine reading comprehension,” in *Proceedings of the 24th Conference on Computational Natural Language Learning*, Online, 2020.

[30] I. Kuzovkin, “Understanding information processing in human brain by interpreting machine learning models,” *arXiv [q-bio.NC]*, 17-Oct-2020.

[31] T.-L. Wong, H. Xie, W. Lam, and F. L. Wang, “A learning framework for information block search based on probabilistic graphical models and Fisher Kernel,” *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 9, pp. 1473–1487, Sep. 2018.