

# A Review of Explainable Artificial Intelligence: Taxonomies, Challenges, Implementation Frameworks and Future Directions

Hussaini Aliyu Idris, Adamu Muhammad, Buhari Aliyu, Sabiu Usman Suleiman

Dept. of Computer Science, Jigawa State Polytechnic Dutse, PMB 7001 Dutse, Nigeria.

Dept. of Electrical and Electronics Engineering Technology, Jigawa State Polytechnic Dutse, Nigeria.

Dept. of Computer Engineering Technology, Bilyaminu Usman Polytechnic Hadejia, Nigeria.

**ABSTRACT:** The recent surge of interest in eXplainable Artificial Intelligence (XAI) within the research community has led to an increasing number of publications in this field. This interest stems from the critical need for interpretability of machine learning models, particularly in sectors such as healthcare and autonomous driving. Consequently, there is an urgent requirement for an updated review to keep up with the current trends in XAI. However, existing literature reviews pose challenges for newcomers, especially those without a computer science background, as they struggle to comprehend complex taxonomies while lacking information about available libraries and frameworks for practical implementation. In this paper, we aim to address these limitations by presenting a simplified taxonomy of XAI that is accessible to readers with diverse backgrounds. By breaking down complex concepts into more understandable components, we provide a comprehensive overview of XAI and its implementation libraries. Moreover, we discuss the challenges faced by XAI, including data privacy and security concerns, the complexity of implementation, bias and fairness issues, and the integration of AI systems with existing infrastructure and finally, we suggest future directions for XAI research, emphasizing the need for responsible AI development, interactive machine learning, AI-assisted education, and further advancements in healthcare applications.

**Keywords:** artificial intelligence, machine learning, neural network, LIME, SHAP, black box, explainable AI, XAI, taxonomy

## INTRODUCTION

The proliferation of artificial intelligence (AI) models across diverse sectors has yielded significant

advancements in problem-solving capabilities [1] [2]. Notably, the rapid progress in the field of deep learning and innovative approaches to data utilization have played a pivotal role in this success. However, a notable consequence of this progress is the increasing complexity of AI systems, rendering them incomprehensible even to AI experts. Consequently, these models have been described as "Black Boxes." The term "Black Box" refers to the opaqueness of AI models, wherein the inner workings and decision-making processes are not readily interpretable or explainable. The complexity arises from the intricate layers and connections within deep neural networks, which enable the models to learn and make predictions based on vast amounts of data. While these models have demonstrated remarkable performance in various domains, their lack of transparency presents significant challenges [3].

The opaqueness of AI models poses obstacles in several areas [4]. First, it hampers our ability to comprehend how the models arrive at their predictions or decisions, limiting our understanding of the underlying logic. Consequently, it becomes challenging to identify and rectify potential biases or errors embedded within the models. Moreover, the lack of interpretability raises concerns regarding ethics, accountability, and trustworthiness. Users, stakeholders, and regulatory bodies may be hesitant to adopt AI systems due to the inability to comprehend and scrutinize their decision-making processes.

In safety-critical domains like autonomous driving and healthcare, where machine learning (ML) is extensively applied, understanding the inner workings

of algorithms becomes crucial. This is where Explainable Artificial Intelligence (XAI) emerges as a vital area of research. XAI, also known as interpretable machine learning, aims to provide techniques that enable a better understanding and validation of ML models. By providing explanations or justifications for the models' predictions, XAI seeks to bridge the gap between the complex inner workings of AI systems and our human understanding [5]. The neural network model, characterized by millions of parameters, surpasses human capabilities in complexity, rendering it versatile yet challenging to comprehend. This complexity underscores the necessity for XAI techniques to provide insights into the decision-making processes of neural networks and other intricate ML models.

Figure 1 demonstrates the increasing attention bestowed upon XAI by the research community in recent years. This growing interest stems from the demand for interpretability of ML models, particularly neural networks, across diverse applications. The need for transparency and interpretability in ML models can even be seen as a component of user experience (UX), extending beyond the purview of data scientists to end-users who seek explanations for decision-making processes.

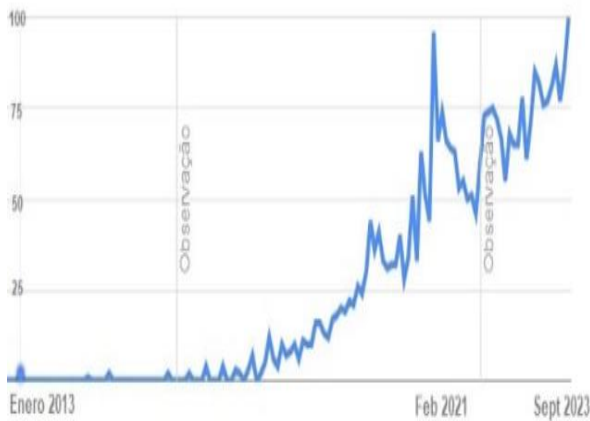


Figure 1 eXplainable Artificial Intelligence Trends (source: "Google Trends")

When constructing a model from data, a trade-off often exists between accuracy and interpretability. Simpler linear models, such as linear regression, offer high interpretability but may struggle to address complex problems effectively. On the other hand, complex models like neural networks exhibit strong generalization capabilities for complex problems but possess low interpretability, as depicted in

Figure 2

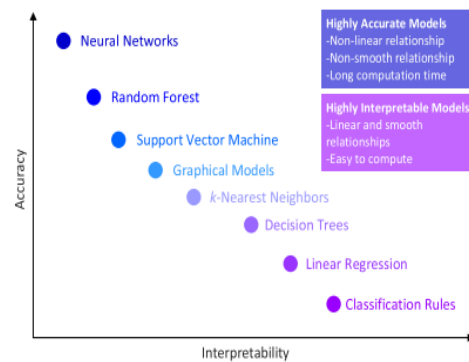


Figure 2  
Trade-Off

Between Accuracy and Interpretability of ML Models [6]

## RELATED WORK

The field of Explainable Artificial Intelligence (XAI) has emerged as a critical area of research, aiming to enhance the transparency and interpretability of AI models for users. This literature review provides an overview of the taxonomy, applications, challenges, and future directions of XAI.

To begin with, The author in [7] conducted a literature using semantic scholar to collect the most recent papers in eXplainable Artificial Intelligence (XAI) using common keywords in multidisciplinary fashion citing the multidisciplinary nature of XAI as reason. Their findings show a significant of adoption of XAI in different sectors in recent years.

Further, the authors in [6] presented the employability of machine learning in 5G and beyond networks where they suggest a shift to explainable AI for future directions. The authors also presented the trade-off between machine learning model's accuracy and its interpretability where the high accuracy of neural networks and random forest as well as their complexity in terms of interpretability are acknowledged.

In addition, the authors in [8] conducted scientific literature review on the taxonomies of eXplainable AI where they introduced a database and a decision tree to help scholar from different background to find the fitting methods for XAI models.

Moreover, the authors in [9] conducted a comprehensive survey on the recent advances of XAI, the effectiveness of the explanation methods. They also delve into the security concerns of XAI. Likewise the authors in [10] explored the opportunities that can be harnessed from the field of XAI and suggest future directions.

Furthermore, the taxonomies, case studies and the lessons learned from XAI are covered by the authors

in [11] where they also suggest the future research directions. In a similar survey by the authors in [12], the conceptual frameworks and the explanation methods are discussed.

Moving forward, the work presented in [13] classifies eXplainable Artificial intelligence (XAI) based on the model's output into image, text, tabular (numerical) and graph. They went on to discuss about the limitation of each classification. While the authors in [14] discuss about the theorem, applications of counterfactuals and causability in explainable artificial intelligence.

Finally, the authors in [15] contributed to the body of knowledge in the XAI field by conducting the review stressing on the taxonomies, application and challenges. They also discussed about the evaluation metrics for XAI and reached the consensus that no specific evaluation is found in the literature for XAI models.

### TAXONOMY OF XAI

In this section, the taxonomy of eXplainable Artificial intelligence (XAI) adopted in this paper is presented as shown in Figure 3. The taxonomy of explanation methods based on the agnosticity or applicability, scope, data type and explanation type as shown in **Error! Reference source not found.** are also discussed.

#### Taxonomy of eXplainable Artificial Intelligence Approach

The field of explainable artificial intelligence (XAI) has garnered significant attention from researchers across diverse backgrounds. This paper aims to contribute to the existing body of knowledge by categorizing XAI into two distinct categories, as visually depicted in Figure 2.

The categorization of XAI techniques plays a pivotal role in understanding the various approaches employed to enhance the interpretability and transparency of AI models. By classifying XAI into different categories, researchers can gain a comprehensive overview of the available methods and their respective strengths and limitations.

Figure 2 serves as a visual representation of this categorization, providing a clear and concise illustration of the two main branches of XAI. The categorization framework adopted in this paper offers a structured approach to organizing and analyzing the diverse range of XAI techniques.

By delving into the specific categories of XAI, researchers and practitioners can gain a deeper understanding of the underlying principles and methodologies employed within each category. This

classification serves as a foundation for further exploration and investigation, enabling researchers to identify gaps in the current XAI landscape and propose novel approaches to address them.

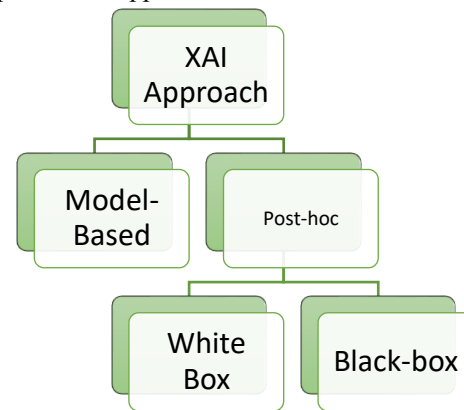


Figure 3 Taxonomy of Approaches for Building XAI Model

The comprehensive categorization of XAI techniques presented in this paper contributes to the ongoing discourse in the field by providing a systematic framework for understanding the diverse approaches to enhancing interpretability in AI models. It serves as a valuable resource for researchers, practitioners, and stakeholders invested in the development and deployment of transparent and explainable AI systems.

There are two main approaches in the field of explainable AI (XAI) for achieving interpretability: the model-based approach, also known as the ante-hoc approach, and the post-hoc approach.

### THE MODEL-BASED APPROACH

The model-based approach focuses on building interpretable machine learning (ML) models, often referred to as "glass box" models. These models are designed with transparency and interpretability as key objectives from the outset. By employing simpler and more transparent algorithms, such as linear regression or decision trees, these models offer clear insights into the decision-making process. The interpretability of these models stems from their explicit rules or decision paths, making it easier to understand how they arrive at predictions or decisions.

#### The Post-hoc Approach

On the other hand, the post-hoc approach involves deriving explanations from complex ML models, which are often considered "black box" models. In certain cases, we may not have direct access to the internal workings of these models, including weights and gradients. This limits our ability to extract explicit

rules or decision paths. In such scenarios, explanations are derived based on the relationship between the

model's input and output. This approach is referred to as the **black box approach**. Despite the lack of direct access to internal model information, various techniques, such as feature importance analysis or local surrogate models, can still provide insights into the model's behavior and decision-making process.

In contrast, when we have access to the internal information of the ML model, such as gradients and weights, the post-hoc approach becomes a **white box approach**. With this level of access, explanations can be derived by analyzing the internal workings of the model, such as feature attribution or sensitivity analysis. The white box approach offers more detailed and fine-grained explanations, as it leverages the model's internal information to provide transparency into its decision-making process.

Both the black box and white box approaches within the post-hoc category aim to address the interpretability challenge posed by complex ML models. While the black box approach focuses on leveraging input-output relationships to derive explanations, the white box approach utilizes internal model information to provide more detailed insights.

#### TAXONOMY OF XAI EXPLANATION METHODS

In this subsection, we present a categorization of methods used in eXplainable Artificial Intelligence (XAI) for deriving explanations. This categorization is based on various factors, including agnosticity or applicability, scope, data type, and the nature of the explanation itself. Figure 4 provides a visual representation of this categorization.

#### EXPLANATION METHOD BASED ON AGNOSTICITY

The first factor considered in this categorization is agnosticity or applicability. Some XAI methods are designed to be model-agnostic, meaning they can be applied to any machine learning model regardless of its specific architecture or type. These model-agnostic methods focus on understanding the model's behavior and decision-making process without relying on internal model information.

On the other hand, there are XAI methods that are model-specific, meaning they are tailored to a particular type of machine learning model or architecture. These model-specific methods leverage the unique characteristics and properties of the model to derive explanations. They may utilize internal model information, such as gradients or weights, to

provide detailed insights into the model's decision-making process.

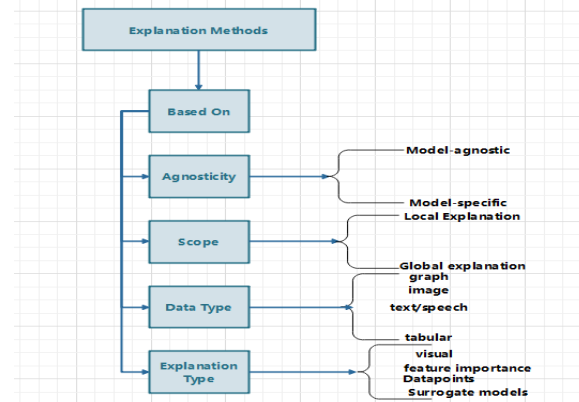


Figure 4 Taxonomy of Explanations Methods for XAI

#### EXPLANATION METHOD BASED ON SCOPE

The scope of the explanation is another important factor in this categorization. Some XAI methods aim to provide global explanations, which offer insights into the model's behavior across the entire dataset or input space. These explanations focus on understanding the overall trends, patterns, and biases present in the model's decision-making process.

In contrast, there are XAI methods that focus on providing local explanations. These explanations zoom in on specific instances or inputs and aim to understand how the model arrived at a particular prediction or decision for those instances. Local explanations are particularly useful for understanding the model's behavior on individual cases and can offer valuable insights into potential biases or errors.

#### EXPLANATION METHOD BASED ON DATA TYPE

The data type of the output of the model's explanation is also taken into account in this categorization. XAI methods can be categorized based on the type of input data they handle. For example, some methods are specifically designed for tabular data, where the input features are structured in a tabular format. Other methods are tailored for handling image data, text data, time series data, or other specific data types. These data-specific XAI methods leverage the unique characteristics of each data type to derive meaningful explanations.

#### EXPLANATION METHOD BASED ON EXPLANATION TYPE

Finally, the nature of the explanation itself is considered in this categorization. Some XAI methods focus on generating feature importance scores,

indicating the relative importance or contribution of each input feature to the model's predictions. Other

methods provide rule-based explanations, offering human-readable rules or decision paths that describe how the model makes decisions. Additionally, there are XAI methods that generate contrastive explanations, highlighting the differences in the model's behavior between different instances or classes.

By considering these factors, the categorization of XAI methods provides a comprehensive framework for understanding and analyzing the different approaches used to derive explanations in the field. This categorization enables researchers and practitioners to explore the various methods available and select the most suitable ones based on the specific requirements of their applications.

#### LIBRARIES, TOOLS AND FRAMEWORKS FOR XAI

In this section, we will explore the various libraries, tools, and frameworks that can be leveraged to implement explainable Artificial Intelligence (XAI) in a programmatic manner. To the best of our knowledge, this work is unique in its focus on implementation tools and libraries within the XAI domain. The primary focus of this exploration will be on tools and libraries available in the Python programming language. Python is widely acknowledged for its popularity within the data science and machine learning communities. Additionally, Python offers a rich ecosystem of well-documented libraries and extensive support forums, making it an ideal choice for implementing XAI techniques.

There are several notable libraries and tools that can be utilized for XAI implementation in Python. These include:

1. LIME (Local Interpretable Model-Agnostic Explanations) is a Pythonic tool that is not specific to any particular machine learning model and is used to provide explanations for the predictions of such models. It works by creating local explanations through an approximation of the model's behavior around the specific instance being explained. LIME is proposed by the authors in [16] in the 2016 and offers variety of libraries for different data types such as tabular, and image data.

2. SHAP (SHapley Additive exPlanations) is a framework designed to interpret the output of any machine learning model. It offers a unified measure of

feature importance and generates global explanations by calculating the contribution of each feature to the prediction. SHAP proposed by the authors in [17] offers various flavor of libraries such as DeepSHAP for neural networks, TreeSHAP for tree-based models, KernelSHAP, and LinearSHAP.

3. ELI5 (Explain Like I'm 5) is a Python library that offers a straightforward and intuitive method for explaining the predictions of machine learning models. It is compatible with various models and provides both local and global explanations.

4. InterpretML is another Python library developed by Microsoft that provides a range of tools for interpreting machine learning models. Similar to ELI5, it supports various models and offers both local and global explanations, along with features for assessing feature importance, generating partial dependence plots, and providing counterfactual explanations.

5. TensorFlow Lattice is a library within the TensorFlow framework that is designed for constructing interpretable machine learning models. It offers a collection of pre-built lattice models that can be utilized for tasks such as regression, classification, and ranking. Additionally, it supports the creation of custom lattice models and provides tools for visualizing the behavior of the model. The library enables the injection of domain knowledge into the learning process through the use of Keras layers that can satisfy constraints such as monotonicity, convexity, and pairwise trust. Furthermore, it provides easy-to-use canned estimators for common use cases, allowing for the incorporation of domain knowledge to enhance extrapolation into different parts of the input space. The library's flexibility and interpretability make it a valuable tool for developing machine learning models that are not only accurate but also transparent and aligned with domain-specific knowledge.

These are just a few examples of the many tools and libraries available for implementing XAI techniques in Python. Each library has its own unique set of features and capabilities, catering to different XAI requirements and scenarios.

By utilizing these Python libraries and tools, developers and researchers can effectively implement XAI methods to enhance the interpretability and transparency of their machine learning models. The availability of well-documented libraries and support forums in Python further facilitates the implementation process, enabling users to leverage the collective knowledge and expertise of the community.

## CHALLENGES OF XAI

To begin with, data privacy and security concerns pose significant challenges in the realm of eXplainable Artificial Intelligence (XAI). As AI systems frequently handle sensitive data, apprehensions about the privacy and security of that data naturally arise. Protecting the confidentiality and integrity of this information becomes paramount, as any breaches could have severe consequences [18].

Further, complexity involved in developing and maintaining AI systems presents another challenge for XAI adoption. Building robust AI models demands substantial investments of both time and expertise. This complexity can deter smaller organizations and individuals from embracing XAI due to the resources required for implementation. As a result, accessibility to XAI techniques becomes limited, hindering the widespread adoption and benefits of transparent and interpretable AI.

Moving forward, bias and fairness represent yet another significant challenge in the context of XAI. AI models are heavily influenced by the data they are trained on, potentially leading to biased outcomes or unfair treatment of certain groups. Addressing these biases and ensuring fairness in the decision-making process of AI systems is crucial for building ethical and trustworthy AI models. Tackling this challenge involves not only collecting diverse and representative training data but also developing techniques to detect and mitigate biases in the decision-making process.

Finally, Integrating AI systems with existing infrastructure or software poses a unique set of challenges. Incorporating AI models into pre-existing simulation environments or software frameworks requires careful consideration of compatibility, scalability, and performance. Ensuring seamless integration without disrupting existing workflows and functionalities can be a complex task that demands expertise in both AI and system integration.

To overcome these challenges, collaborative efforts between researchers, policymakers, and industry professionals are essential. Robust data privacy and security frameworks need to be established to protect sensitive information while enabling the transparent analysis of AI models. Simplifying the implementation process and providing comprehensive resources for XAI adoption can encourage smaller organizations and individuals to embrace interpretable AI. Additionally, continuous research and development in bias detection and mitigation techniques can help address fairness concerns. Lastly, close collaboration between AI experts and system integrators is vital to seamlessly incorporate AI

systems into existing infrastructures while minimizing disruption.

## CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we present a structured taxonomy of explainable artificial intelligence with the aim of serving as an overview and reference to beginners, researchers and practitioners in the surging field of XAI research. The paper provides some useful tools, frameworks and libraries for implementation of XAI programmatically which serves as unique contribution to the existing body of knowledge in XAI field.

Furthermore, as the field of Artificial Intelligence (AI) continues to progress, future research and development in eXplainable AI (XAI) will likely focus on several key areas:

1. Responsible AI: With the growing societal impact of AI, ensuring the development and deployment of AI systems in a responsible manner becomes crucial. Future research in XAI will prioritize the promotion of fairness, transparency, and accountability in AI systems. This includes addressing biases, ensuring ethical decision-making, and establishing regulatory frameworks to govern the use of AI technologies.
2. Interactive machine learning: To enhance the usability and effectiveness of AI systems, future research in XAI will delve into developing methods that enable users to interact with AI models. These methods will allow users to provide feedback, correct errors, and actively participate in the learning process of AI models. By incorporating user feedback, AI models can adapt and improve their performance over time, leading to more accurate and reliable outcomes.
3. AI-assisted education: The application of AI in education has shown great potential in personalized learning experiences and educational tools. Future research in XAI will continue to explore and innovate in this area, leveraging AI to provide tailored and adaptive learning experiences for students. This includes developing intelligent tutoring systems, automated feedback mechanisms, and AI-driven educational content generation, ultimately enhancing the effectiveness and accessibility of education.
4. AI in healthcare: The healthcare industry has witnessed significant advancements in the application of AI, and this trend is expected to continue in the future. XAI research will focus on expanding the use of AI in healthcare, particularly in areas such as medical imaging analysis, clinical decision-making support systems, and personalized treatment plans. By leveraging AI, healthcare professionals can benefit from improved diagnostics, more accurate prognoses,

and optimized treatment strategies, ultimately enhancing patient care and outcomes.

These areas of research reflect the evolving needs and possibilities within the AI and XAI domains. By focusing on responsible AI, interactive machine learning, AI-assisted education, and AI in healthcare, researchers and practitioners can contribute to the development of AI systems that are not only powerful and efficient but also transparent, accountable, and beneficial to society as a whole.

#### REFERENCES

- [1] A. Guha, "Building Explainable and Interpretable model for Diabetes Risk Prediction," *Int. J. Eng. Res. Technol.*, vol. 09, no. 09, 2020, doi: 10.17577/IJERTV9IS090510.
- [2] S. M. Lundberg et al., "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nat. Biomed. Eng.*, vol. 2, no. 10, p. 749, 2018.
- [3] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 2522–5839, 2020.
- [4] D. S. A. K. Gattani Tanuj Subhash, "Artificial Intelligence Approaches to Uncover Cyber Security," *Int. J. Eng. Res. Technol.*, vol. 12, no. 08, 2023, doi: 10.17577/IJERTV12IS080006.
- [5] H. Mankodiya, M. S. Obaidat, R. Gupta, and S. Tanwar, "XAI-AV: Explainable Artificial Intelligence for Trust Management in Autonomous Vehicles," in *2021 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, 2021, pp. 1–5. doi: 10.1109/CCCI52664.2021.9583190.
- [6] M. E. Morocho-Cayamcela, H. Lee, and W. Lim, "Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions," *IEEE Access*, vol. 7, pp. 137184–137206, 2019, doi: 10.1109/ACCESS.2019.2942390.
- [7] A. Jacovi, "Trends in Explainable AI (XAI) Literature." 2023.
- [8] T. Speith, "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 2239–2250. doi: 10.1145/3531146.3534639.
- [9] A. Rawal, J. McCoy, D. B. Rawat, B. M. Sadler, and R. St. Amant, "Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives," *IEEE Trans. Artif. Intell.*, vol. 3, no. 6, pp. 852–866, 2022, doi: 10.1109/TAI.2021.3133846.
- [10] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: <https://doi.org/10.1016/j.inffus.2019.12.012>.
- [11] M. I. Khalid A. Eldrandaly Mohamed Abdel-Basset and N. M. Abdel-Aziz, "Explainable and secure artificial intelligence: taxonomy, cases of study, learned lessons, challenges and future directions," *Enterp. Inf. Syst.*, vol. 17, no. 9, p. 2098537, 2023, doi: 10.1080/17517575.2022.2098537.
- [12] G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts," *Data Min. Knowl. Discov.*, 2023, doi: 10.1007/s10618-022-00867-8.
- [13] G. Vilone and L. Longo, "Classification of Explainable Artificial Intelligence Methods through Their Output Formats," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 3, pp. 615–661, 2021, doi: 10.3390/make3030032.
- [14] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," *Inf. Fusion*, vol. 81, pp. 59–83, 2022, doi: <https://doi.org/10.1016/j.inffus.2021.11.003>.
- [15] W. Ding, M. Abdel-Basset, H. Hawash, and A. M. Ali, "Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey," *Inf. Sci. (Ny)*, vol. 615, pp. 238–292, 2022, doi: <https://doi.org/10.1016/j.ins.2022.10.013>.
- [16] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should {I} Trust You?: Explaining the Predictions of Any Classifier," *CoRR*, vol. abs/1602.04938, 2016.
- [17] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.
- [18] I. OGREZEANU et al., "Privacy-Preserving and Explainable AI in Industrial Applications," *Appl. Sci.*, vol. 12, no. 13, p. 6395, 2022.

