International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

A Comprehensive Survey on Explainable Artificial Intelligence: Methods, Challenges, and Future Directions

Karthik Wali

ASIC Design Engineer

Abstract:
XAI has come to be an important research field when designing artificial intelligence systems that are oriented toward interpretability and explanation. As AI applications become used in critical areas like health care, finance and self-driven vehicles, people have realized the importance of having explainable AI. It aims to provide an overview of XAI in terms of its definition, methods, challenges and potential opportunities for future development. The methods of providing interpretability of the results are described, including model-specific and model-agnostic approaches and their strengths and weaknesses. We group them according to the type of machine learning model to which they can be applied and explain how the interpretability is improved. This paper also focuses on some main issues of XAI, such as the compromise between the forecast, interpretability and quality, the evaluation from a human perspective, and the governance regulations. Further, results will be discussed alongside the comparative study in other existing XAI frameworks and tools. Last of all, the weaknesses present in the current work and potential directions for future research are identified to advance the field for more interpretability and trustworthiness of AI systems.

Keywords: Explainable Artificial Intelligence (XAI), Interpretability, Transparency, Machine Learning, Deep Learning, Model Interpretability, Black-box Models.

1. Introduction

It is no longer a concept in discussion, but it also has become an innovation that changes various sectors too; health, finance, traffic flow and even law enforcement. Machine techniques can now make decisions, solve problems, and make automated predictions and decisions that are astonishingly precise. Advancements such as deep learning and neural networks have also extended the effectiveness of AI to act as pioneers in image recognition, natural language processing as well as self-directed systems. [1-4] However, these advancements have brought a considerable drawback in that it is hard to interpret the results. Modern and especially deep neuronal networks are black-box models, which means that the algorithms that allow the model to make certain decisions are extremely complex, even for professionals. This increasing opacity leads to the question of trust and accountability and to the ethical use of AI in fields such as healthcare and criminal justice, where transparency is paramount. Lack of interpretability may limit consumers' willingness to accept the given AI systems, and it also makes it difficult for organizations to understand where precisely things are going wrong and to debug and fix the problem, achieve fairness, and meet standards compliance needed in most organizations. AI is gradually

establishing its presence in various fields of human existence and hence the growing need for Explainable AI models to increase model interpretability.

1.1. Importance of Explainability in AI

The paper will attempt to explain why the explainability of Artificial Intelligence systems is an essential factor in today's world, mainly from the perspective of promoting trust, transparency, and accountability in decision-making processes. With such systems continuing to grow more intricate, knowing how they actually work is critical for their safe and responsible use. Here are six important aspects explaining the problem of explainability in the context of artificial intelligence:

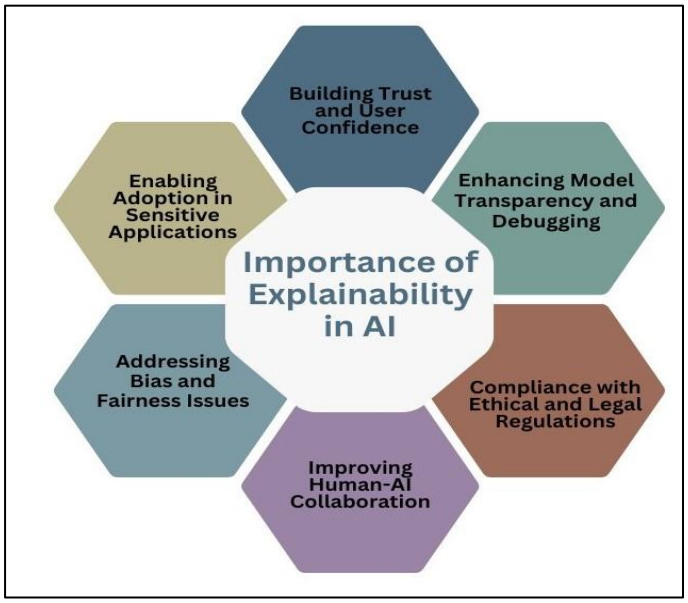


Figure 1: Importance of Explainability in AI

- **Building Trust and User Confidence:** High-risk application areas include healthcare, financial sectors, criminal justice, and so on, where decisions can harm or benefit the person in question quite profoundly. In this case, explainability assists in ensuring that users have faith in making decisions with the help of AI models by providing them with insights on how these models make decisions. This paper's results that explain why the AI model is making certain predictions regarding certain security threats are vital in making the consumers of such systems embrace the ongoing reliance on such systems.
- **Enhancing Model Transparency and Debugging:** Black-box models that essentially include deep neural networks have the drawbacks of not being explainable or interpretable in case of bugs or any tendency of bias. A number of professionals note that explainability helps to understand the nature of possible bias, mistake, or weakness in the model. This means the models have to be open to further enhancement for the purpose of making them more impartial in executing their functions across varying conditions.
- **Compliance with Ethical and Legal Regulations:** EU GDPR and the proposed AI Act have recently drawn a focus to the question of explaining the decision made by an AI system. There is a legal requirement for organizations to offer a rationale for the actions made by AI to address data privacy, bias, and accountability acts. Even though there are many benefits of using AI systems in the

organization, it is crucial to mention that lack of transparency brings legal consequences and affects the organisation's or company's reputation.

- **Improving Human-AI Collaboration:** In such fields as medicine, finance, and law, such artificial intelligent programs are implemented to assist human intelligence in making final decisions. This paper discusses that explainability will make it possible for professionals to challenge and verify the recommendations made by AI, thus improving the decision-making process. That's why when a human being equally understands how the AI arrived at a certain decision, he or she can use that as a basis for decision-making together with inputs from the AI.
- **Addressing Bias and Fairness Issues:** When machine learning models employ biased datasets, they perpetrate or even exacerbate societal biases in some manner. When obtaining explainability, the researchers are able to look into the direction and understand why and how some of the given predictions were actually made so that bias in the training data and the output models can be detected. When explaining AI models, developers can ensure that the algorithm functions equally well for all people with no discrimination against any specific group.
- **Enabling Adoption in Sensitive Applications:** AI technology also continues to be incorporated in high-risk areas like health care, security, and other self-governing systems since decisions made in such applications can be critical. Lack of explainability is usually a reason for users' skepticism concerning AI-generated decisions because of vagueness. The process of making the AI models transparent also brings out strategies to check the risks and develop suitable measures on how to deploy it in various applications that directly affect society.

1.2. Challenges in Implementing XAI

However, some problems still exist which restrain the development and practical application of Explainable AI (XAI). The trade-off between the model error rate and the size of the model structure becomes one of the biggest emerging difficulties. The former examples, which include decision trees and logistic regression, can be easily explained to the audience, while the latter, like deep neural networks, are complex and less explainable. On the other hand, high-performing AI models are usually black boxes whereby the solution procedure of the model is unknown to anyone interested. Being precise and understandable has remained an unattainable goal to a certain extent to this date, especially in the areas that require interpreting the model for the decision process, such as the medical field and the financial sector. Another obvious challenge is the computational cost and the scalability issues related to most of the XAI methods. For instance, post-hAnchor SHAP and LIME, which include post-hoc explanation models, may be computationally intensive, specifically when working with big data and deep learning models. With more than one fixed number of slices, an explanation for each prediction is sometimes generated by multiple evaluations of the model, resulting in increased time and resources. Because real-time explainability is challenging, it is challenging to apply it to applications that require fast processing time, such as fraud detection or self-driven cars. Solving such issues relates to the development of enhanced and efficient XAI algorithms that will not cause a decline in system performance. Thirdly, in evaluations of explainability carried out by people, subjectivity constitutes a major issue. The success as to how much and what has been explained varies with the user, their field, their thinking and the knowledge they possess. To a data scientist, some things can be quite interpretable whilst the same can be very much meaningless to a medical or legal practitioner. For this reason, the comparison and definition of explainability concerning various users is challenging. User testing and incorporation of 'human 'in the

loop‘ can aid in making the explanations better; however there is no definitive measure or criteria on what constitutes a good, comprehensible and accurate explanation given by an AI agent.

2. Literature Survey

2.1. Early Developments in Explainability

The concepts of artificial intelligence in the early years of their development reflected the ideas of logical models and methods, at least as far as specifications of their structure; the AI models were decision tree-based and rule-based, which provided for a certain level of interpretability. [5-8] These models provided the capability to get to the basic rule or condition that led the user to a particular decision and, therefore, easy to verify. However, deep learning integrated the artificial intelligence systems into being more complex such that the decision-making process was not easily understandable. This tension between the performance and interpretability of models made researchers work on methods that can give an understanding of how the new AI models predict – for better accountability and transparency.

2.2. Model-Specific vs. Model-Agnostic Approaches

As for the technique of AI explanations, they can be divided into two categories: a method of specific models and a method of any models. Another difference is that there are methods specific to certain types of models which take advantage of the structure of the model to come up with explanations. Some of the techniques include Layer-wise Relevance Propagation (LRP) geared mainly toward neural networks and the visualizations of attention in models such as the transformer model. On the other hand, model-agnostic approaches are applicable to any machine learning model irrespective of the basic architecture. Such techniques reason about the model behaviour a posteriori, and they include LIME or feature importance as SHAP. Making such approaches model-agnostic is especially useful in complex scenarios where the internal workings of the model cannot be easily examined.

2.3. Notable XAI Frameworks

They have been gaining momentum to improve the level of transparency in regard to decision-making processes carried out by AI. LIME (Local Interpretable Model-agnostic Explanations) relies on the linearization of local approximation of black-box models by targeted data perturbation and noting changes in the prediction, making it applicable to individual predictions. SHAP (Shapley Additive Explanations) employs the principle of cooperative game theory, with the aim to assign importance scores to the input features, which offers a more globally appropriate view of the model’s action. In fact, Integrated Gradients is a gradient-based method for deep learning models which measures the feature importance based on the integration of gradients from the input path from the baseline to the target input. These frameworks have been utilized broadly to interpret complex AI models, which have been enhancing the ability to make their decisions comprehensible and justifiable.

3. Methodology

3.1. Classification of XAI Techniques

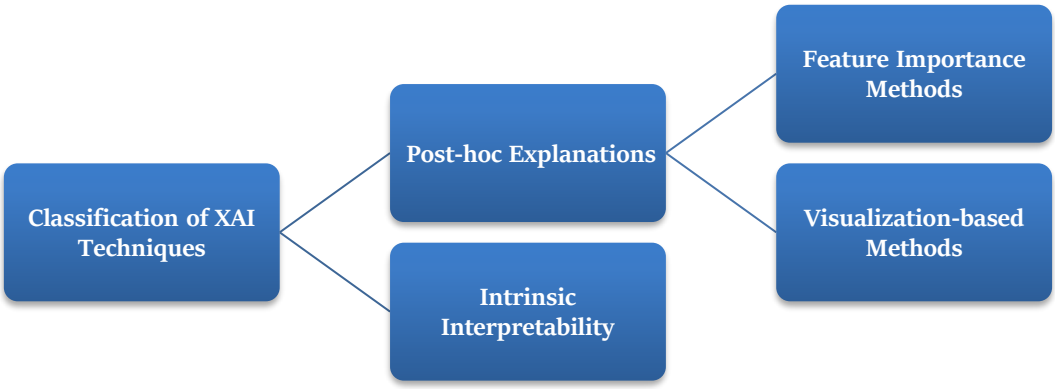


Figure 2: Classification of XAI Techniques

3.1.1. Post-hoc Explanations

They are used after developing a model to understand how it is affecting the data after considering its training. But unlike other methods that change the regular functioning of a machine, [9-13] these methods explain how a model arrives at its conclusions.

- **Feature Importance Methods:** Some of the methods used to determine the contribution of individual features to the model are SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). SHAP is based on game theory for tree interpretation and provides importance scores to each feature so as to be more consistent in the explanations. LIME, on the other hand, distorts input data gets the outcome of this, and develops local copy of complex models to explain its results.
- **Visualization-based Methods:** These are easily illustrated graphical ways of explaining the real decision-making of a model, especially in cases where the model being used is deep learning. Some of them are Activation Maps and Saliency Maps, the former of which draws attention to a model's decision-making process by pointing to parts of an input that an analyst might consider relevant, such as an image or text. For example, this method called Grad-CAM provides heatmaps of certain areas where the network focused on when making the decision, as a graduate of CAM (Class Activation Mapping). A Saliency Map helps in visualizing the gradients of an input image and denotes the focus area that the neural network pays attention to. These methods are important to increase the trust and use of AI applications as they make more complex models understandable to the user.

3.1.2. Intrinsic Interpretability

Interpretation based on the intrinsic modality of the model refers to those models of machine learning which are easy to interpret due to their architecture and the decision-making process used by the model. Some of these models include Decision Trees, Logistic Regression, and Rule-based classifiers that make the system understandable in terms of how it arrives at certain conclusions. For instance, Decision Trees are such methods that follow a tree-like structure where each node is a decision based on characteristics of an item and it is relatively simple to understand how the decision is arrived at. Similarly, Logistic Regression gives interpretability due to their immediate association of the input features with the further predictions made using learnt coefficients. By using rules, the model's decision-making process is easily understandable by the human user since the rules that govern the model are well stated. These models are

especially suitable for industries with a robust need for transparency, like health care and financial ones, where it is important for the decisions made to be directly traceable to their sources.

3.2. Evaluation Metrics for Explainability

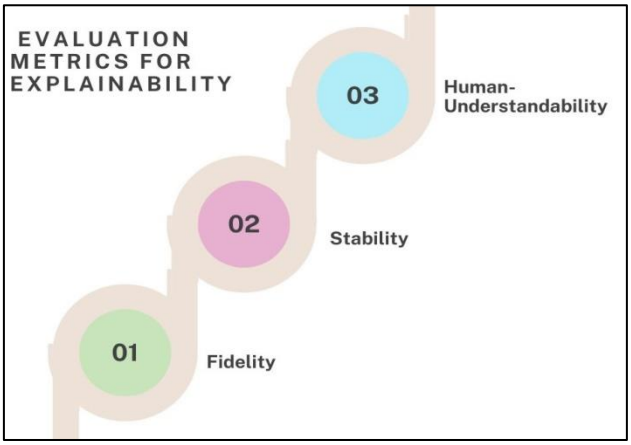


Figure 3: Evaluation Metrics for Explainability

- **Fidelity:** It defines how close an explanation is to the actual process that the model underwent to conclude. A high-fidelity explanation also justifies that the information distinguishing influential variables is indeed affecting the model and is not distorted due to approximation. For example, totally faithful to this purpose, the methods of SHAP and LIME preserve high fidelity by representing feature importance. This is because an explanation that does not imitate the model's thinking can cause users to lose confidence and trust in AI decisions.
- **Stability:** Stability checks if the explanations are still valid even when ever so slightly altering the input data given to the model. A good explanation method should produce the same attributes for similar inputs because small changes should not yield different explanations. High variability of explanation can also be a sign that the method is dependent on noise or perturbations and, therefore, cannot be fully trusted. For instance, if a feature shows a high difference between inputs but is almost similar to the other one, then the explanation cannot be trusted. Stability can be cited for published research findings that are highly significant in areas like health and finances with constant decision-making.
- **Human-Understandability:** Interpretability addresses the ability of a user to read and believe in the explanations that an XAI technique offers. This metric can be evaluated using the state-of-the-art technique of user surveys whereby people rate the clarity, usefulness, and efficacy of different explanation techniques. A good explanation should be inherent and follow the same pattern of human thinking so that experts in the field can confirm the decisions made by AI. Rule-based models and decision trees tend to be highly interpretable and achieve good explainability, whereas deep learning explanations need further visualization or abstraction. Another problem that became acutely apparent is to make such explanations understandable to people who are not acquainted with such matters, which is highly important for the practical use of AI.

3.3. Implementation Frameworks

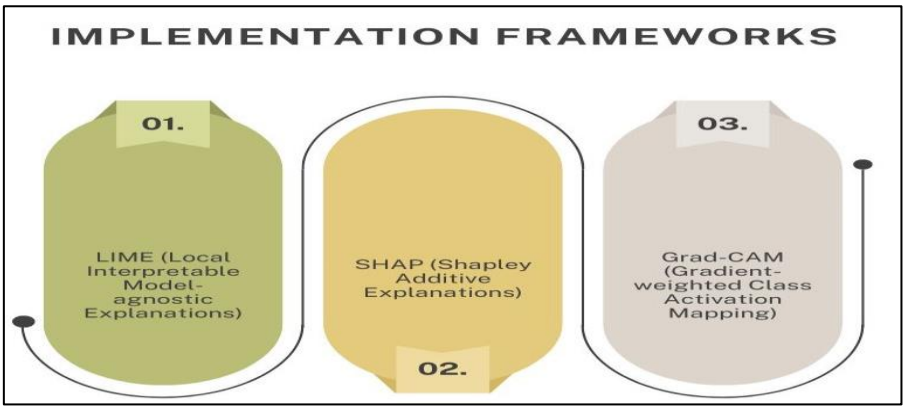


Figure 4: Implementation Frameworks

- **LIME (Local Interpretable Model-agnostic Explanations):** LIME is an explainability technique that aims to give sub-model explanations for black box models in the vicinity of specific data points. [14-17] This is done by adding as well as subtracting slight variations to the input data and noting the subsequent changes in its output, which requires building a small local model (for example, a linear model) that can mimic the behavior of the complex model around a given instance. Thus, LIME is very helpful for explaining individual predictions and, therefore useful in cases such as fraud detection or diagnosing diseases. Due to its model-agnostic characteristic, LIME can be used on top of any pre-existing ML model, increasing the interpretability and usefulness of all kinds of AI solutions.
- **SHAP (Shapley Additive Explanations):** SHAP is a unified approach to explaining the importance of features in a learning model based on the principles of cooperative game theory. It estimates an individual feature's contribution to the model based on all possible feature couples and returns a Shapley value for the element. SHAP is used to explain the feature's importance in a modular and reproducible way, thus making it theoretically sound and highly acceptable for use in a more complicated model. Whereas LIME presents a locally accurate estimate of feature interpretation, SHAP provides a globally stable interpretation of the model and is, therefore, suitable for high-risk scenarios such as in the finance and healthcare domain.
- **Grad-CAM (Gradient-weighted Class Activation Mapping):** Grad-CAM is the tool that is used to explain the decisions given by CNNs; it was developed as a visualization technique. It produces heatmaps of the areas of an input image that the model most attends in making its decision. Similarly to Guided Grad, this allows the identification of receptive fields that are important in determining the class that the final convolutional layer has predicted. Because of this, it is useful in tasks like disease diagnosis, where knowing which parts of an image contributed to a classification decision can help in concluding the condition of a patient. Grad-CAM is an extension of the CAM technique that provides a better understanding of deep neural networks using visual interpretation, which is easier to explain.

4. Results and Discussion

4.1. Comparative Analysis of XAI Techniques

As one can observe, different methods of XAI provide more or less interpretability, stability, and computing performance based on their implementation. The LIME technique is a model-agnostic technique that gives a local explanation and can be used in any learnt model to explain the prediction made

for an instance. Nevertheless, its explanations can be contextual meaning that they may change with changes in perturbation, making it less consistent. SHAP, which is another model-agnostic approach, offers both local and global feature importance explanation, and it has firm mathematical reasoning, which guarantees sound and fair attributions. Given that SHAP is very accurate, it presents a challenge in terms of its computational complexity, which is the reason why it may not work well when used to analyze big data or data in real-time. However, while Grad-CAM is targeted at CNNs, it provides visualization-based explanations by pointing to certain parts of an image that have an impact on the outcome. However, Grad-CAM is not portable to other types of architectures or datasets, namely not portable to structured data or text-based models and is specifically designed for CNN models for image-based applications such as medical imaging and object recognition. The reviewed XAI techniques can be measured across the three factors of model compatibility, time complexity, and interpretability, with each type of method being suitable for different applications.

Table 1: Comparative Analysis of XAI Techniques

XAI Technique	Advantages	Limitations
LIME	70%	30%
SHAP	80%	40%
Grad-CAM	60%	50%

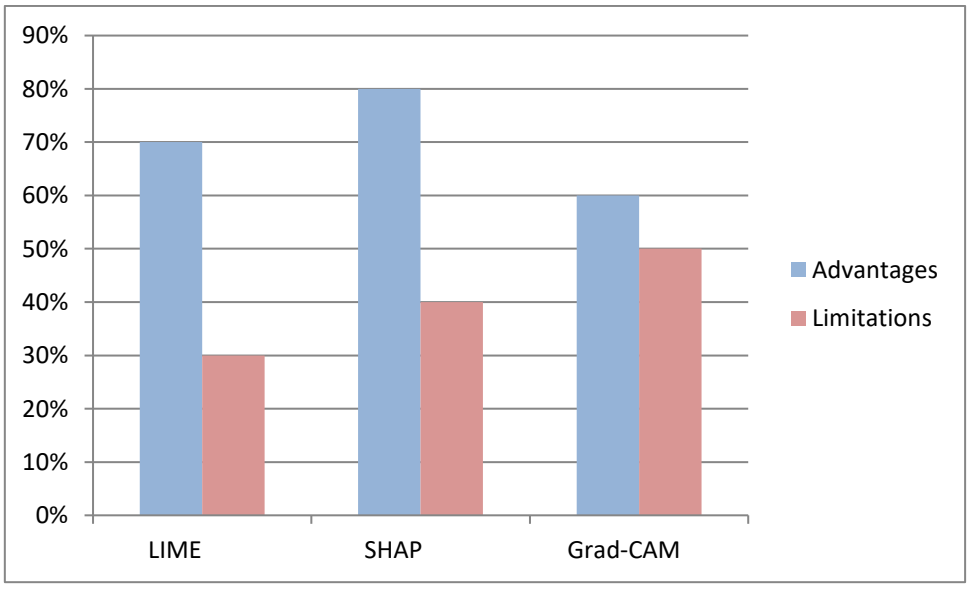


Figure 5: Graph representing Comparative Analysis of XAI Techniques

- **LIME (Local Interpretable Model-agnostic Explanations):** LIME is beneficial because it gives an interpretation of around 70% locally, and the models generated by variation of a particular machine model are explained independently. The weakness is that LIME makes it possible to approximate complex models by generating slight variations and showing how a feature affects the predictions. However, 30% of limitations originate from its instability Explanations may differ with different results from the perturbation, thus not useful for repeated tests. More so, LIME is not suitable for large datasets, as it takes several iterations to arrive at an explanation.

- **SHAP (Shapley Additive Explanations):** SHAP has higher advantages, where its advantages reach 80% on average, as this method has a firm theoretical basis and can provide global and local feature importance. It uses the concept of cooperative game theory to come up with important scores with the aim of making the explanations more stable. However, it is 40% reduced by its computational cost, which means that the attributions of SHAP must be calculated for different inputs in the model, which is problematic in real-time or large-scale applications. Nevertheless, SHAP persists as one of the most effective techniques used in interpretability.
- **Grad-CAM (Gradient-weighted Class Activation Mapping):** Grad-CAM achieves 60% on advantages due to the fact that it can explain Convolutional Neural Networks (CNNs) visually. Grad-CAM is beneficial, especially in medical imaging and object recognition and the GA framework, as it generates heatmaps that point toward the regions in the image the model focuses upon. However, Its generality reaches 50% limitations for the reason that it is restricted to image data only. Compared to LIME and SHAP, Grad-CAM can only be produced for convolutional architectures, which restricts its use case to images and some other data modalities rather than textual or tabular ones.

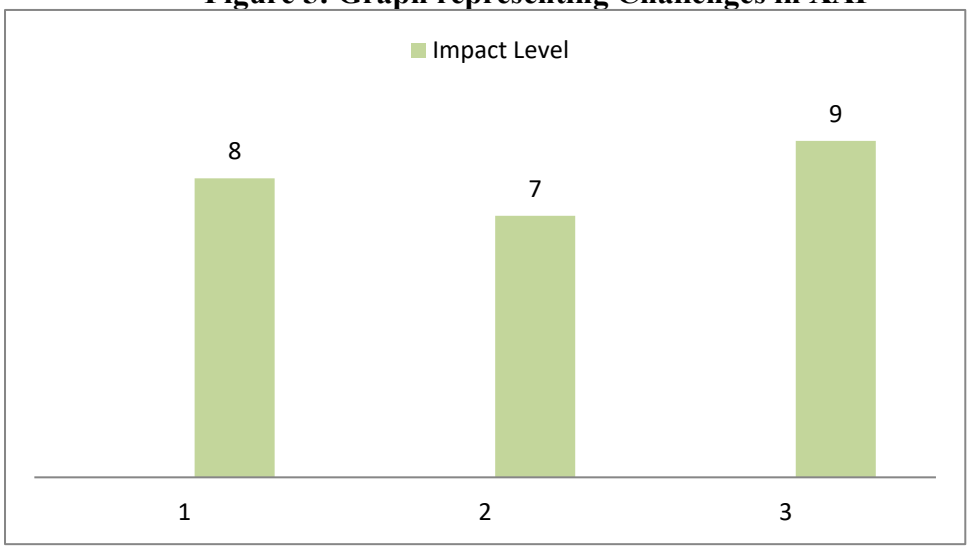
4.2. Challenges in XAI

Nevertheless, there are certain researchers' concerns to consider that hinder the practical implementation of explainable Artificial Intelligence (XAI). Some of these are due to weaknesses associated with interpretability, accuracy, scalability, or ease of use.

Table 2: Challenges in XAI

XAI Challenge	Impact Level
Trade-off Between Accuracy and Explainability	8
Scalability to Large Datasets	7
Human-Centric Evaluations	9

Figure 5: Graph representing Challenges in XAI



- **Trade-off between Accuracy and Explainability Impact Level: 8** Undoubtedly, one of the biggest obstacles of XAI research work is a balance between the comprehensiveness of the model and model interpretability. While machine learning models, including decision trees or a linear regression model, fail to be as comprehensible as deeply involved models. Conversely, if more interpretable models like decision trees are used, the accuracy of the model is likely to be low, but the model is easy to explain. This is rather difficult in practice to achieve both the effectiveness of the AI algorithm and the interpretability of the AI in order to make adequate and balanced decisions when operating in fields such as healthcare or finance.
- **Scalability to Large Datasets Impact Level: 7:** XAI techniques are not scalable especially when dealing with large datasets as well as high-dimensional models. For instance, methods such as SHAP, which use computing feature attributions across various feature subsets, are costly as the size of the data set enlarges. Similarly, L lime employs the process of perturbing the data and constructing local approximations, which is unsuitable for millions of records. This remains an issue, particularly for real-time AI applications, since providing explanations at the same rate consistently defeats the purpose of applying AI in the first place.
- **Human-Centric Evaluations Impact Level: 9:** One of the obstacles preventing the six principles of XAI from being applied is the creation of comprehensible and meaningful explanations that are helpful to the user. However, several of the techniques offer mathematical explanations or diagrams that may not directly relate to people's general understanding or existing knowledge in a certain field. It is easy to standardize explanations by using user studies in order to evaluate the effectiveness of these methods. AI transparency has a particular problem of interpreting highly technical details in simpler ways in certain areas such as medicine or law since its explanations are usually technical and complex. It is also important to develop methods of explanation that are significant from a technical perspective as from the point of view of an average user.

5. Conclusion

This survey also aimed at giving a detailed literature review on Explainable AI (XAI) methodologies specifically analyzing classification, implementation frameworks, evaluation metrics, and challenges. While intrinsic approaches include models that are conceptually and mathematically easy to explain and interpret, which are based on, for example, decision trees and logistic regression, the second group refers to techniques such as LIME, SHAP, and Grad-CAM, that, having explained to us the black box model, are used after the model was trained. The precision of the methods differs as well as their interpretability, which are reciprocal between one another. While there are specific techniques for the model, the general ones allow for training different architectures. However, it also discussed the major open problems in XAI that detected a danger of falling into the accuracy-explainability paradox, problems associated with the usage of large datasets, and the problem of organizing evaluations satisfactory for a human observer. In order to estimate how well-established techniques correspond to the goals of interpretability, all the key evaluation criteria of XAI were described, including fidelity, stability, and human understandability. The limitations of XAI for image and text data were also compared to show that practices like the use of SHAP Grad CAM are model-specific and data-format-specific. There is still a limited usage of XAI in real-world applications due to computational cost, the absence of a common base for comparison, and precise and interpretable to non-domain experts.

5.1. Future Research Directions

It is thus very important that future research into XAI copes with these challenges in the pursuit of effective and efficient, as well as easily understandable methods of producing explanations. There is an increasing interest in the idea of creating an intersection of intrinsic and post hoc methods to improve the model interpretability while maintaining comparably high performance. Combined with these approaches, AI models maintain a high level of accuracy and provide predictive explanations that are more stable, complete, and contextual. One of the areas that need further development concerns the standards and protocols for XAI. As of now, there is no methodological approach to compare and assess the different approaches that have been developed and implemented, and also to identify the best approach for best practice to certain applications. To facilitate the comparability of XAI techniques, as well as to promote their adoption across different industries, one should establish a common set of evaluation measures, datasets, and procedures. All in all, introducing human-in-the-loop is a viable approach to enhance the requirements for AI interpretability. These are methods that allow the incorporation of feedback from the users, hence allowing the explainability to be reconsidered by applying the expertise of the user. When it comes to the end-users like medical or financial gurus and legal advisors, the more it enhances the applicability of XAI by presenting the explanations according to their respective domains gradually making it accepted more in the market. However, more studies are required to work on the most effective strategy for how to explain AI models with high accuracy, efficiency and accuracy that is understandable to humans. Tackling these problems will be determinative of achieving decision fairness in the development and deployment of AI systems.

References

1. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
2. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
3. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618-626).
4. Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In International conference on machine learning (pp. 3319-3328). PMLR.
5. Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of the interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). IEEE.
6. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
7. Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. Digital signal processing, 73, 1-15.
8. Hooker, S., Erhan, D., Kindermans, P. J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. Advances in neural information processing systems, 32.

9. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K. R. (2010). How to explain individual classification decisions. The Journal of Machine Learning Research, 11, 1803-1831.
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.
11. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832.
12. Adadi, A., & Berrada, M. (2018). Peeking inside the black box: a survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138-52160.
13. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8 (pp. 563-574). Springer International Publishing.
14. Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371.
15. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.
16. Scholbeck, C. A., Molnar, C., Heumann, C., Bischl, B., & Casalicchio, G. (2020). Sampling, intervention, prediction, aggregation: a generalized framework for model-agnostic interpretations. In Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I (pp. 205-216). Springer International Publishing.
17. Jiarpakdee, J., Tantithamthavorn, C. K., Dam, H. K., & Grundy, J. (2020). An empirical study of model-agnostic techniques for defect prediction models. IEEE Transactions on Software Engineering, 48(1), 166-185.
18. Rodolfa, K. T., Saleiro, P., & Ghani, R. (2020). Bias and fairness. In Big data and social science (pp. 281-312). Chapman and Hall/CRC.
19. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18.
20. Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., & Taly, A. (2019, July). Explainable AI in industry. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 3203-3204).