International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

"EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR RELIABLE DECISION-MAKING IN AUTOMATED SYSTEMS"

KIRANKUMAR PUNDLIK MOHURLE
Lecturer
Computer Science & Engineering Department
Somayya Diploma In Engineering
kiranmohurle0446@gmail.com

SHWETA ARUN SAHARE
Lecturer
Computer Science & Engineering Department
Somayya Diploma In Engineering
shwetasahare1212@gmail.com

YUGANT RUPESH DHOKE
Student
Computer Science & Engineering Department
Somayya Diploma In Engineering
yugantdhote014@gmail.com

POOJA RAVI UPPLETI
Student
Computer Science & Engineering Department
Somayya Diploma In Engineering
upplitesujatha@gmail.com

Abstract: In recent years, AI is being leveraged for different applications by automated systems to assist in making decisions on various fields like healthcare, financial sector, cybersecurity, transportation, manufacturing, and public services etc. While contemporary machine learning models can make accurate predictions, several of the sophisticated models are “black boxes” and users are unable to comprehend the very reason behind certain decisions. This opacity can hinder transparency, pose challenges in identifying mistakes or bias, and restrict the appropriate applications of AI in critical areas. Explainable Artificial Intelligence (XAI) is one of the solutions to this problem that provides tools to make AI predictions, recommendations, and decision processes more understandable to humans. This paper discusses the contribution of XAI for increasing the reliability of the automated decision-making systems. It provides information on the key components of transparency, interpretability, and accountability, fairness, error detection, and user trust. Additionally, the study investigates a variety of explainability methods such as feature importance, local explanations, visual interpretation, rule-based interpretations, and model independent approaches. It has been reported that explainability can aid in the improvement of AI decisions and the preparation of better human oversight, where organizations can detect

inappropriate or ambiguous AI decisions that can lead to significant consequences. Concurrently, the paper notes that there are issues such as the quality of explanations, difficulty of computation, privacy, security, and trade-offs among model quality, interpretability, and model complexity. The study suggests that explainable AI is a crucial factor in building intelligent, reliable, accountable, transparent, and effective AI-powered systems that can be deployed in realistic environments for decision making.

Keywords: Explainable AI, Automated Systems, Decision-Making, Interpretability, Reliability, Transparency

Introduction:

AI is an integral part of contemporary computer systems. Data analysis, identification of complex patterns, prediction of future events, or making decision making has been made possible by AI based technologies which can process vast amounts of data and analyze complicated patterns with limited human input. Automated systems including fraud detection, medical diagnosis, credit assessment, autonomous vehicles, recommendation systems, industrial monitoring and cyber security are all widely using the machine learning and deep learning models. As such, there has been a lot of contributing to the speed and efficiency of many of the decision-making processes which make the use of these technologies more popular.

The adoption of AI has brought forth significant questions about the accuracy and transparency of AI-driven decision-making, however. Some of the more sophisticated machine-learning algorithms, such as deep neural networks and ensemble-based solutions, feature intricate internal architectures that are hard for a typical person or even for a technical adept—to decipher. An information system can offer a prediction or suggestion without stating explicitly which are important factors in the prediction. These kinds of systems are often referred to as ‘black-box’ systems.

When automatic decision making has a direct and important impact on an individual or organisation or critical operations a lack of clear explanations can be a significant issue. They might, for instance, want to find out why an AI system has rejected a financial application, declared a patient to be at high risk, detected a cybersecurity threat, or advised some operational action. It's hard to know if it's right, fair, biased and/or based on bad information.

The pursuit of Explanation (xAI) is one way to overcome these limitations and it has become a critical field of research. XAI is concerned with techniques and methods that facilitate understanding of the behavior and output of an application of artificial intelligence by human users. XAI does not just aim for high accuracy of prediction, but to explain actual outcomes and elicit meaning from which input factors played the greatest part.

Another benefit of explainability is that it can help in the reliability of automated systems, as it can facilitate the identification of errors and the supervision by humans. When users grasp why a prediction was made, they can discern instances where the AI system might have relied on irrelevant features, biased data, or dubious patterns. Explanations can thus help verify AI outputs more effectively and reduce risk of uncritically blindly accepting incorrect automated decisions.

Another key factor of XAI is also accountability. An automated decision-making process could demand an explanation of the workings of the system and show that the decisions are based on appropriate information. Explainable systems can produce more transparency and enable developer(s), auditor(s), domain expert(s) or end user(s) to assess the decision-making process better.

Several methods are proposed to enhance the explainability of AI. There are some techniques based on inherently interpretable models like decision trees or rule-based techniques, and others which deliver explanation of complex models after prediction. When it comes to feature importance, the values of the variables whose inclusion has the greatest impact on a model are revealed, while local explanation approaches give explanations for individual predictions. Explain relationships between input data and system decisions with visualization and example-based explanations can also assist users.

While explainable AI has its benefits, it also presents a number of difficulties. An increasing number of models can be very interpretable but offer poor predictive accuracy at times. Depending on user, explanations can get too technical or too simple. Additionally, you could have privacy or security issues if you divulge too much information about a model. As a result, it is necessary to have a balance of accuracy, transparency, interpretability and security when developing reliable automated systems.

Hence, it becomes crucial to study how to obtain interpretable and explainable AI for the future development of trustworthy automated systems. AI's ability to deliver insightful explanations alongside high-quality AI capabilities empowers companies to create systems that assist rather than replace human judgment. In the end, XAI can help to develop more responsible, reliable, transparent, and human-centered applications of AI.

Related Works:

Explainable Artificial Intelligence has emerged as an important research area because of the growing use of complex machine learning and deep learning models in automated decision-making systems. Although advanced AI models can provide highly accurate predictions, their internal decision processes are often difficult to understand. This lack of transparency has encouraged researchers to develop methods that can explain how AI systems arrive at particular outputs and how such explanations can improve reliability, trust, accountability and human supervision.

Ribeiro, Singh and Guestrin (2016) made an important contribution to the field through the development of a model-independent explanation technique for individual predictions. Their work introduced the idea that users should not only receive a prediction from a machine learning model but should also be able to understand the factors responsible for that prediction. Their approach demonstrated that local explanations can help users examine the behavior of complex classifiers and determine whether model decisions are reasonable. This work became an important foundation for later research on interpretable machine learning and explainable AI.

Lundberg and Lee (2017) proposed a unified framework for interpreting model predictions by measuring the contribution of individual features to a particular output. Their approach provided a systematic method for explaining predictions across different machine learning models. The work was particularly significant because it connected several existing feature attribution methods within a common theoretical framework. Their contribution strengthened the use of feature-based explanations for understanding complex AI models and later became widely applied in practical explainable AI systems.

Samek et al. (2017) focused on the evaluation of visualization techniques used to understand what deep neural networks have learned. Their study highlighted the importance of identifying which parts of the input data are most influential in producing a particular model prediction. The authors demonstrated that visual explanations can provide useful insights into the internal behavior of neural networks. Their findings were especially relevant to applications involving image recognition and other high-dimensional data where direct interpretation of model parameters is difficult.

Adadi and Berrada (2018) presented a comprehensive survey of Explainable Artificial Intelligence and examined the growing need to make black-box models more transparent. They discussed different categories of explanation methods and highlighted the relationship between interpretability, transparency and trust. Their study emphasized that explainability is increasingly important when artificial intelligence is used in critical applications where decisions can have significant consequences. They also identified major challenges associated with understanding complex machine learning systems and highlighted the need for further research in evaluating explanation quality.

Guidotti et al. (2018) conducted a detailed survey of methods developed for explaining black-box models. The authors classified explanation approaches according to the type of model being explained, the scope of the explanation, and the form in which explanations are presented. Their work showed that explainability can be achieved through several techniques, including feature importance, rule extraction, local explanations, prototypes and visualization. The study demonstrated that there is no single explanation method suitable for every problem and that the selection of an appropriate XAI method depends on the application, model architecture, and user requirements.

Miller (2019) examined explanation in artificial intelligence from the perspective of social science research. He argued that explanations generated by AI systems should be designed according to the way people naturally understand and interpret decisions. The study emphasized that technically accurate explanations may not always be meaningful to ordinary users. Miller highlighted that effective explanations should be selective, contrastive, and understandable within the context of the user's expectations. This work expanded the study of explainability beyond technical methods and introduced important human-centered considerations into XAI research.

Barredo Arrieta et al. (2020) provided an extensive review of explainable artificial intelligence by discussing its concepts, classifications, opportunities, and challenges. Their study connected explainability with broader principles of responsible artificial intelligence, including fairness, accountability, transparency, privacy, and trust. The authors proposed a structured taxonomy of XAI methods and discussed how explainability can be incorporated into different stages of the AI lifecycle. Their work emphasized that reliable AI systems

should not only achieve high predictive accuracy but should also provide understandable and justifiable decisions.

Linardatos, Papastefanopoulos, and Kotsiantis (2021) reviewed a wide range of machine learning interpretability techniques and examined their applicability to different model types. The authors discussed both model-specific and model-agnostic methods and explained how interpretability can be achieved through visualization, feature attribution, surrogate models, and rule-based techniques. Their study showed that the rapid development of machine learning models has increased the need for methods that can explain model behavior without significantly reducing predictive performance.

Vilone and Longo (2021) examined different concepts of explainability and the approaches used to evaluate explainable artificial intelligence systems. Their study highlighted that explainability is not a single measurable property but includes several dimensions, such as understandability, transparency, completeness, and usefulness. The authors pointed out that one of the major challenges in XAI research is the absence of standardized evaluation methods. They emphasized that explanations should be assessed not only according to technical correctness but also according to whether users can understand and effectively use them.

Nauta et al. (2023) conducted a systematic review focusing on methods used to evaluate explainable AI. Their research demonstrated that although many explanation techniques have been proposed, the evaluation of these techniques remains inconsistent across studies. The authors identified a need for more quantitative, systematic, and reproducible methods for measuring explanation quality. Their work highlighted important evaluation dimensions such as correctness, completeness, consistency, robustness, and user understanding. The study contributed to the growing recognition that explainability should itself be evaluated before an XAI system can be considered reliable.

Van Mourik et al. (2024) provided a tertiary review of the current state of explainable artificial intelligence. Their study examined findings from existing surveys and reviews to identify broader patterns in XAI research. The authors observed that explainability research has expanded considerably across multiple domains, but several challenges continue to remain. These include the lack of common definitions, limited standardization of evaluation criteria, differences in user requirements, and difficulties in comparing explanation techniques across applications. Their work indicates that the field is gradually moving from simply developing new explanation methods toward evaluating their reliability, usability, and practical effectiveness.

Overall, the reviewed literature demonstrates that Explainable Artificial Intelligence has evolved from a narrow focus on interpreting individual machine learning predictions to a broader research area concerned with transparency, trust, fairness, accountability, human understanding, and responsible automated decision-making. Earlier studies mainly concentrated on developing explanation techniques such as local explanations, feature attribution, and visualization, while more recent research has increasingly focused on the quality, evaluation, and practical usefulness of explanations.

The literature also reveals that different XAI methods provide different advantages and limitations. Techniques such as LIME and SHAP are useful for explaining complex black-

box models, while visual and rule-based approaches can make model behavior more understandable to human users. However, there is still no universal explanation technique or standardized evaluation framework that can be applied to every automated decision-making system.

A major research gap identified from the existing literature is the need to establish a stronger connection between explainability and the overall reliability of automated systems. Many studies focus either on the development of explanation techniques or on their interpretability, but comparatively less attention is given to how explanations can improve error detection, decision verification, accountability, human supervision, and operational reliability simultaneously. Therefore, the present study focuses on examining Explainable Artificial Intelligence as a mechanism for supporting reliable decision-making in automated systems by considering both technical interpretability and practical decision reliability.

Objectives of the Study:

1. To examine the role of Explainable Artificial Intelligence in improving the transparency and reliability of automated decision-making systems.
2. To analyze major XAI techniques used for interpreting and explaining the predictions generated by artificial intelligence and machine learning models.
3. To identify the major challenges and benefits associated with the implementation of explainable AI in real-world automated systems.

Material and Methods:

This study takes a descriptive and analytical research methodology to analyze the use of XAI in enhancing automated system reliable decision-making. The study relies primarily on secondary data sources derived from existing published research papers, technical reports, conference proceedings, books and studies on the topics of 'artificial intelligence', 'machine learning', 'automated decision systems', 'interpretability', 'transparency', and 'trustworthy AI'.

Among the topics the methodological framework covers is the understanding of how various explainability techniques can help in the interpretation of complex machine learning systems. Special focus is given to some of the more popular XAI methods, including SHAP, LIME, feature-importance methods, decision tree, rule-based explanations, and visualization-based techniques. These methods were analyzed based on their accuracy in explaining individual prediction, important identified variables, and giving understandable information to the user.

Different environments of automated decision-making such as healthcare, financial services, cyber security, industrial automation, transportation and recommendation systems are also taken into consideration in the study. These applications were chosen because the AI-generated decisions in these systems can impact users, organizations, and processes.

A comparative method was used for the analysis. The interpretability and transparency of MAI techniques were compared with these and were followed by a computational study of the techniques for determining the suitability for various models and suitability for reliable

decision-making. The study also reviewed relevant reliability issues including fairness, fairness accountability, human supervision, error identification, trust and decision consistency.

As the study is conceptual and literary, personal information and information gathered from direct human participants is not and will not be collected. The findings were systematically developed based on the examination, comparison, and interpretation of existing knowledge related to AI explainability and automated decision-making systems.

Results and Discussion:

The result analysis shows that explainability can be crucial in increasing the trustworthiness of AI-driven automated systems. The more users understand not only what a particular automated system outputs but also the key factors that go into this output, the more useful the automated system is. XAI methods can deliver varying levels of explanation, depending on the model architecture and decision environment.

Table 1. Comparison of Major Explainable AI Techniques

XAI Technique	Main Function	Interpretability Level	Major Advantage	Limitation
LIME	Explains individual predictions using a simpler local model	High	Easy to understand for specific predictions	Explanations may vary between similar instances
SHAP	Measures the contribution of each feature to a prediction	High	Provides detailed and consistent feature importance	Can require high computational resources
Feature Importance	Identifies the most influential input variables	Medium to High	Simple and useful for understanding model behavior	May not explain individual predictions clearly
Decision Trees	Represents decisions through rules and branches	Very High	Naturally interpretable and easy to visualize	Performance may decline with very complex data
Rule-Based Explanation	Converts decisions into understandable logical rules	High	Suitable for human interpretation	Difficult to generate for highly complex models
Visualization Methods	Presents model behavior through graphs, maps, or plots	Medium to High	Improves understanding of patterns and relationships	Interpretation may depend on user expertise

The comparison shows that no single explainability technique is suitable for every automated system. Decision trees and rule-based techniques provide strong interpretability, while SHAP and LIME are useful for explaining complex black-box models. The choice of technique

therefore depends on system complexity, application requirements, and the type of explanation required by users.

Table 2. Influence of XAI on Reliability Factors in Automated Systems

Reliability Factor	Without Explainability	With Explainability	Expected Impact
Transparency	Low	High	Users can understand why decisions are generated
Error Detection	Difficult	Easier	Incorrect patterns and unusual decisions can be identified
Accountability	Limited	Improved	Decision responsibility can be examined more effectively
User Trust	Moderate or Low	Higher	Users gain confidence when decisions are understandable
Bias Identification	Difficult	More visible	Unfair feature influence can be detected
Human Supervision	Limited	Stronger	Human experts can verify AI recommendations
Decision Reliability	Uncertain	Improved	Explanations support validation of system outputs

The analysis presented in Table 2 suggests that explainability influences several dimensions of system reliability. Transparency and human supervision are particularly important because they allow decision-makers to examine whether AI-generated outputs are based on reasonable and relevant information. Explainability also supports accountability by creating greater visibility into automated decision processes.

Table 3. Application of Explainable AI in Different Automated Systems

Application Area	Automated Decision	Role of XAI	Reliability Benefit
Healthcare	Disease prediction and risk classification	Explains symptoms and clinical factors influencing predictions	Supports medical verification and safer decisions
Banking and Finance	Credit approval and fraud detection	Identifies financial features influencing decisions	Improves fairness and transparency
Cybersecurity	Threat and anomaly detection	Explains why activities are classified as suspicious	Helps security experts verify alerts
Autonomous Transportation	Object recognition and driving decisions	Shows factors influencing system actions	Improves safety and human understanding
Manufacturing	Fault detection and predictive maintenance	Identifies causes of equipment failure predictions	Reduces unexpected operational failures
Recommendation Systems	Product, content, or service recommendations	Explains why an item is recommended	Improves user confidence and personalization

The application-based analysis here shows that explainability can serve multiple purposes in various scenarios and contexts for different domains. For the healthcare industry, explanations can help healthcare professionals confirm predictions generated by AI. They can aid in levelling the playing field regarding the lending process in the finance industry. In such settings as cybersecurity and industrial sectors, explanations may be vital for figuring those threats out that are real and those that are in error.

The results of the study suggest that XAI has the potential to greatly enhance the accuracy of automated decision-making systems by enhancing clarity and validation of AI-generated results. Unlike standard AI systems, which are based on accuracy, precision, recall or speed of processing, trusted AI decisions must focus on other more important qualities like transparency, fairness, accountability and level of understandability to humans.

The comparative analysis of the techniques of XAI demonstrates that other methods offer different types of explanation. Many of the techniques, like LIME and SHAP can be especially helpful in understanding highly complex machine learning models when they cannot be directly understood. These techniques enable users to investigate the impact of each feature and the "how" or "why" of why a given prediction was made. Decision trees and rule induced systems, however, give explanations by means of natural understandable decision structures.

The findings also indicate explainability could be useful to enhance error detection. Bad data, bad training set, irrelevant variables and unknown runtime conditions are some scenarios where the automated systems can produce wrong predictions. In a full black-box model it is hard to know if such problems exist. Explainability brings extra information to support anomaly investigations by developers or domain experts to check if the model uses the correct reasoning.

One more important discovery relates to the level of trust among users. When people are offered reasons for their decisions and if that reason is done through AI, they are more likely to accept those decisions. But trust cannot be established with only an explanation, however. Explanation should also be correct, meaningful, and related to the user. Descending to fallacies in explanations could lead to over confidence in inaccurate forecasts.

The study also emphasizes the criticality of explainability to detect bias. When the training data sets incorporate an inequity or inequality, the automated system that is programmed to use them may perpetuate the same inequity or inequality. Explainability methods can help uncover if sensitive or inappropriate factors are unduly affecting system decisions. This is useful for the developer to explore model behaviour and to strengthen the fairness of the model.

However, even with the latest AI, human oversight continues to be crucial in areas like healthcare, banking, transportation, and cybersecurity, where sophisticated algorithms might not capture all interactions or responses. Despite the power of AI, complex environments like healthcare, banking, transportation, and cyber security rely on humans for tasks that may not be fully automated by AI, such as interpretive decisions or complex interactions. Results suggest that the introduction of XAI in the domain may enhance the efficacy of human-AI collaboration by enabling decision making by domain experts based on AI suggestions.

Rather than artificial intelligence itself being the decision-making power, explainability can be used to make it a transparent decision-support system.

However, there are a number of drawbacks. There are some highly accurate models that are very complex, and it is easy to simplify their behaviour so that it makes human readable sense and end up with an incomplete interpretation. Unfortunately, some computations methods might become complex when applied to the very large real-time data. Besides, users can request diverse types of explanations. A technical explanation may not be suitable for a Medical Professional, Bank Customer, or a System User in general.

From the results, it appears therefore that explainable AI efforts should be built following the requirements of the application and the users. Too much accuracy is not enough to understand reliability. Such an automated decision system should possess a high level of predictive acuity, should be easy to explain, should have appropriate levels of human oversight, and should have accurate error detection and bias detection capabilities.

The overall findings of the study show that XAI has a crucial contribution to the development of responsible automated systems. Explainability makes AI more humanlike and transparent, empowering users to interpret, challenge and confirm AI-generated decisions. Implementing it can improve the safety, justice, accountability and trustworthiness of automated decision-making systems.

Conclusion:

Explainable Artificial Intelligence (XAI) is now a key method of enhancing the trustworthiness, transparency, and accountability of automated decision-making systems. The use of AI is now becoming more and more prevalent, and to ensure that users understand how and why certain businesses make certain decisions, it will be essential for them to understand why and how.

The findings of the study demonstrate that some of the XAI techniques like visualization, decision trees, LIME, rule-based explanations, and feature importance could help simplify complex ML models. These methods facilitate identification of key decision factors, error checking, consideration of possible bias, and checking the reasonableness of AI-generated results.

Further, the analysis shows how explainability can enhance human supervision, and boost confidence in automated systems. Nevertheless, the efficiency of XAI relies on the quality and clearness of the explanations given, as well as the context of the explanations. The models can still be very complex making it hard to fully understand them; and there can be methods of explainability that need extra computing power.

The result is that the ability to make reliable decisions which are automated should not be judged solely on the basis of predictive accuracy of automated systems. It ought to also take into account transparency, equity, accountability, interpretability, and human oversight. Therefore, this can help to build more trustworthy, responsible, and dependable AI systems in the real world when integrating explainable AI into automated systems.

References:

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

2. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

3. 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, 51*(5), Article 93. https://doi.org/10.1145/3236009.

4. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18. https://doi.org/10.3390/e23010018.

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

6. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007.

7. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys, 55*(13s), Article 295. https://doi.org/10.1145/3583558.

8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778.

9. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K. R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE*

Transactions on Neural Networks and Learning Systems, 28(11), 2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820.

10. van Mourik, F., Jutte, A., Berendse, S. E., Bukhsh, F. A., & Ahmed, F. (2024). Tertiary review on explainable artificial intelligence: Where do we stand? *Machine Learning and Knowledge Extraction, 6*(3), 1997–2017. https://doi.org/10.3390/make6030098.

11. Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion, 76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009.