WCRG Global Research
WORLD COMET RESEARCH GROUP

International Journal of Modern Innovations and Emerging Trends
Vol. 6, No. 1, 2023
Doi: 10.67228/30715628/IJMIET-2023PI1J5W
PP. 01-11

Original Article

The Emergence of Explainable Artificial Intelligence in Modern Decision Systems

Dr. Meena Krishnan

Assistant Professor, Anna University, India.

Received: 12-03-2022 Revised: 12-22-2022 Accepted: 01-01-2023 Published: 01-03-2023

ABSTRACT

Through the application of Artificial intelligence (AI), there has been the establishment of the technology as a cornerstone in current decision systems in essential fields like health, finance, transport, industrial automation, and the governance of citizens. Although traditional AI and machine learning frameworks, specifically deep learning models, have shown outstanding predictive performance, their nature is not enlightened by default, and as such, they have cast considerable doubt on the issues of trust, accountability, fairness, and regulatory compliance. It is thanks to this limitation that Explainable Artificial Intelligence (XAI) as a paradigm has emerged, aimed at ensuring that AI-driven decisions are easy to understand and interpret by the human stakeholders without major performance reduction. This paper is a thorough and stepwise analysis of how XAI was created in current decision systems. It starts with the historical contextualization of the development of AI, as an expert system driven by rules, to a data-driven black-box model and the increasing demand to explain decision-making processes. The paper provides a critical literature review of the current research on XAI methods classifying them into model-intrinsic and post-hoc methods of explanation, and discussing their relevance to various fields. An elaborate methodology is suggested, that incorporates explainability protocols into the AI choice channel, such as information pre-processing, model order, explanation creation and human-centered assessment. Additionally, the paper evaluates the experimental findings and case-based debates on how XAI enhances transparency, end-user trust, compliance with regulations, and system resilience. Popular explainability methods are also compared and evaluated including SHAP, LIME, saliency maps, and rule extraction. The results indicate that explainable models, in addition to increasing interpretability, can also help to improve debugging, bias detection and ethical AI deployment. The paper ends by recommending the current challenges, areas of open research, and future roles of XAI in the development of responsible and human-centered intelligent decision systems.

KEYWORDS

Explainable Artificial Intelligence, Decision Support Systems, Model Interpretability, Transparent AI, Trustworthy AI, Ethical AI, Human-Centered AI, Machine Learning Explainability.

WCRG Global Research
ISSN: 3071-5628
PP: 01-11.

1. INTRODUCTION

1.1. Background

Ever since the commercial beginnings of the so-called AI, the technology has gone through a radical shift in which the simplistic and rule-based systems of reasoning were developed into highly advanced information-driven learning systems. During its early development period, AI studies concentrated on explicitly programmed logic, expert systems and purely symbolic knowledge representations, in which decision making could be transparent, deterministic and easily understandable. These primitive setups allowed the existence of a clear line of reasoning and made it easy to understand and confirm the results made by humans. Scalability, adaptability, and robustness were however hampered by them depending on handcrafted rules and restricted knowledge bases when dealing with complex, uncertain or high-dimensional real world environments. The birth of machine learning brought a considerable paradigm shift in the development of AI through the capability to learn the pattern directly based on the data and not adopting regular rules. This transition was also facilitated by the progress in deep neural network advancement, gain in computational power and access to large scale data sets. Deep learning models performed exceptionally in the fields of image and speech recognition, natural language processing, and autonomous control systems as well. These achievements have not negated the internal decision-making processes of modern AI systems have become more secretive and they tend to be black-box models, which are challenging to comprehend, even to an expert. Lack of transparency in safety critical and regulated applications is a critical issue, where the decisions should be explanationable, audit-able, and legally defensible. Therefore, the development of AI has led to an urgent requirement of the necessity to match predictability with transparency, responsibility, and credibility of contemporary intelligent systems.

1.2. Need For Explainability in Modern Decision Systems

Fig 1 - Need For Explainability in Modern Decision Systems

1.2.1. Transparency and Accountability Requirements

The use of modern decision systems is increasingly important in terms of high stakes in fields like healthcare, finance, law enforcement and autonomous technologies. Under such circumstances, stakeholders should be in a position to comprehend the way and manner in which certain decisions are taken. Explainability is further explained to give transparency of the course of action behind the

ISSN: 3071-5628
PP: 01-11.

model predictions, thereby making accountability possible. In the absence of the clear explanation, it is hard to audit decisions, allocate responsibility, and adhere to the regulatory and ethical standards.

1.2.2. Trust and User Acceptance

The adoption of AI-driven decision systems depends on the user trust as a critical factor. Although black-box models are very accurate, they are usually not accepted without doubts because they are opaque. Explainable AI will help increase trust by only letting users determine whether the decision made in the system complies with the domain knowledge and expectations. As soon as the user is capable of following the arguments of the recommendations, there is a greater chance of adoption and productive use of AI assistance.

1.2.3. Risk Mitigation in High-Stakes Applications

Unexplainable decisions may have severe consequences in any safety-critical and mission-critical environment. Explainability is what facilitates detection of mistakes and biases and aberrant behavior within models before it leads to detrimental results. In real-life applications, XAI helps to deploy and more reliably operate a system through validation and verification processes, potentially making the system safer to deploy.

1.2.4. Regulatory Compliance and Ethical Considerations

A growing regulatory attention such as data protection and algorithmic accountability systems require that automated decisions be interpretable. Fairness, non-discrimination and transparency are some of the ethical AI principles that are based on explainability. The decision systems implemented must comply with the requirements of the society and the law, which can only be achieved by incorporating explainable mechanisms that can guarantee the responsible deployment of AI.

1.3. Artificial Intelligence in Modern Decision Systems

The Artificial Intelligence (AI) is now an essential facilitator of the modern decision-making process that has revolutionized the process of formulation, evaluation, and implementation of complex decisions in a vast array of uses. Through the application of sophisticated algorithms of machine learning, statistical inference, and optimization methods, AI systems can process large amounts of heterogeneous data and yield actionable insights that are beyond the cognitive capacity of human beings. The features of AI in the modern decision-making environment are risk assessment, pattern recognition, predictive analytics, and real-time control, which can enhance efficiency, accuracy, and consistency of results. Artificial intelligence-based decision systems are being applied within industries like healthcare to help clinicians process their medical images, patient data, and diagnostic information so as to aid early disease diagnosis and disease-specific treatment planning. With in the field of finance, AI can be used to generate automated credit scoring, detect fraud, and conduct algorithmic trading based on identifying hidden indicators in transactional data. On the same note, AI supports predictive maintenance, adaptive control, and intelligent navigation in dynamic and act-of-uncertain environments in industrial systems and autonomous systems. Such uses prove the increased dependency on AI as a decision-support mechanism, but not a simple automation tool. Nonetheless, the growing independence and intricacy of AI-based systems of decision making pose serious questions on transparency, accountability and human supervision. Most of the contemporary models, especially the deep learning ones, and the ensemble models are used as black boxes, where it is not easy to determine how the two think. Such opaqueness may impede faith, reduce adoption and be ethically and legally hazardous in sensitive uses. Consequently, explainability, fairness, and robustness have increasingly become the focus of a decision system driven by AI. Through integrating high-performance learning algorithms with fully transparent and explainable decision mechanisms,

ISSN: 3071-5628
PP: 01-11.

contemporary AI systems will be able to promote informed, responsible and human-intelligible to make decisions in an ever-more complex world.

2. LITERATURE SURVEY

2.1. Interpretability Vs. Explainability

Interpretability versus explainability the notion of explainable artificial intelligence (XAI) literature has ubiquitously covered the difference between the similar notions of interpretability and explainability since the two terms discuss various dimensions of model transparency and user comprehension. Generally, interpretability can be defined as how well a human being can directly understand the inner workings and decision making logic of a model without particular external tools being used. Linear regression, logistic regression, decision trees and rule-based systems are the common types of model that can be called interpretable since the parameters, decision paths, and the contributions of the features can be studied easily and interpreted. Explainability, conversely, is often considered to go along with complex or black box models, especially deep neural networks, ensemble procedures, and other high capacity learning systems whose internal representations are not particularly understood or readable. Explainability is about producing post-hoc explanations to justify or rationalize predictions of a model, and not uncover the internal mechanisms of the model. A few studies highlight that explainability is a user- centric and task specific concept, whilst interpretability is a model- centric property. Moreover, according to researchers, the degree and kind of description must be adjusted to the intended audience, including domain competents, regulators, or end users, and to the context of their application, e.g. safety-critical and high-stakes decision environments. This changing differentiation adds to the importance of the flexible explanation implicit structures that minimize the conflicting aims of transparency, usability, and predictive performance.

2.2. Categories Of Explainable AI Methods

Explainable AI approaches can broadly be divided into intrinsic explainability paradigms and post-hoc explainability approaches, and each approach has unique benefits as well as drawbacks. Intrinsic explainability models are meant to be transparent in nature and this means that the user can get the individual interpretation of what was learned by directly interpreting the relation between the input and the output. Some of the examples are decision trees, rule-based classifiers, sparse linear models, and generalized additive models give explicit decision rules or feature level contributions. The models are particularly appropriate to the areas where the emphasis is not on as high a predictive precision as on transparency and accountability. Post-hoc explainability methods, on the other hand, are used after black-box model training and are used to explain the behavior of complex and high performing models. These will involve the feature attribution techniques, sensitivity analysis, and contribution scores; surrogate modeling techniques that give approximations of black-box models using simpler interpretable models and the visualization-based explanations that show internal representations or decision boundaries. Post-hoc approaches are more flexible and can use any trained model but present difficulties on the issue of explanation fidelity, stability and computational overhead. The literature is always keen to point out a trade-off between model accuracy, interpretability, and computational complexity and no single XAI approach is effective across all situations. This means that the choice of a good XAI method should take into account distinct needs of the application such as demands of transparency, real-time demands and trust among users.

2.3. Domain-Specific Xai Applications

Explainable AI has found domain-specific uses indicating the increasing significance of the transparency in a variety of sectors, especially in high-stakes decision-making settings. XAI methods

ISSN: 3071-5628
PP: 01-11.

have been used in healthcare to describe diagnostic predictions, disease-risk estimations and treatment advice, and thus, facilitate clinical decision-making and help physicians have more confidence in AI-assisted systems. Hypotheses in this field are essential in ensuring model conduct is acceptable, finding possible biases, and patient security. Explainability is important in the financial industry in regulatory compliance, credit scoring, fraud detection and fairness audits, where the results of decisions should be explained to regulatory bodies and other interested persons. Clear models and explanations assist the organizations to comply with the legal provisions and allow minimizing the probability of discrimination practices. XAI is used in autonomous and cyber-physical systems to allow human operators to comprehend system behavior, predict system behavior, and take action whenever required, especially in the safety-critical situations. Regardless of the great advancements, the literature outlines the overwhelming difficulties associated with the fidelity of the explanation, human cognitive constraints, and the correspondence between explanations and real model thoughts. These concerns demonstrate why standardized measures of evaluation, user-friendly explanation architecture, and domain-mindful XAI systems are necessary to make sure that explanations do not merely have to be technically accurate, but must be meaningful and actionable to a final user.

3. METHODOLOGY

3.1. Proposed Explainable Decision System Architecture

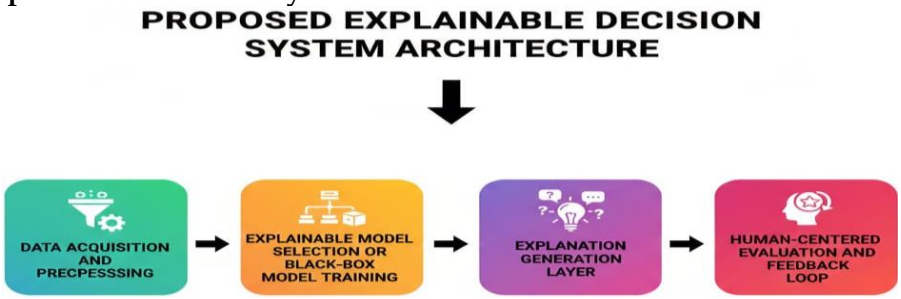

Fig 2 - Proposed Explainable Decision System Architecture

3.1.1. Data Acquisition and Preprocessing

The initial step of the proposed architecture is to do careful data collection using the heterogeneous sources then preprocessing of the data like removal of any repetitive elements in the dataset, cleaning the data, selecting features and treating missing data. This measure will guarantee quality, consistency, and relevance of data that is essential in producing credible model predictions and explanations. It is also important to perform proper preprocessing to limit the propagation of bias as well as improve model transparency through simplifying feature representations. Consequently, the downstream decision-making processes can be explained better.

3.1.2. Explainable Model Selection or Black-Box Model Training

During this step, high-performance black-box models are chosen or inherently interpretable models are chosen depending on the need of the application. The design of interpretable models is transparent and black-box models focus on predictive precision and demand ancillary explanation models. The decision of model depends on the constraints in the domain, riskiness, and the regulatory demands. This step determines the performance/interpretability trade off in the system architecture.

3.1.3. Explanation Generation Layer

The explanation generation layer has the role of generating human understandable insights that justifies model predictions. It uses the right explainability techniques, including feature attribution, surrogate models or extraction of the rules based on the underlying model type. Such explanations

ISSN: 3071-5628
PP: 01-11.

may be produced on a local scale or global scale to facilitate instance-specific or general system knowledge. It is the layer that provides the initial access between complex action of model behavior and human thinking.

3.1.4. Human-Centered Evaluation and Feedback Loop

In the last phase, the human in-the-loop evaluation is included, where the quality of generated explanations, their usefulness, and credibility are evaluated by domain experts or end users. Models, methods of explanation, and data representations are refined through repeated use of feedback gathered through users. It is a recurring feedback loop which improves system reliability, trust from users, and contexts of explanations. Finally, it makes explainability consistent with the real world decision-making requirements.

3.2. Mathematical Foundations of Explainability

It is mathematical that explainability is based on predictive modeling by comprehending the role of individual input features on the ultimate model output. In this case, assume a general predictive function, such that the response variable y is a response of several input features, f of x one, x two, through to x n. The learned mapping of the model denoted by f in this formulation can be as straightforward as a linear relationship to a very nonlinear deep neural network. Explainability aims to break down this complicated role in a way that makes human beings able to think about how variations in each input feature impact the desired result. Technology Feature attribution techniques are a core part of this mathematical construction. The goal of these methods is to estimate the value of every individual feature x i to the total prediction y. This is usually done by finding an importance score, denoted by phi i which is the marginal impact of a feature on the model output. This score can be conceptually defined as the expected model prediction at a given known value of the feature x i, less the expected model prediction over all possible model inputs. Simply put, it is a measure of the extent to which knowledge of a specific feature increases the prediction of the model relative to an average base prediction. When phi i is positive, it means that a feature contributes to the expected value, and conversely, a negative value of the phi i means that the feature suppresses the predictions. It is based on expectation and is a principle-based and model-agnostic approach to explainability, which can work with both interpretable models and black-box systems. The approach to isolate the effect of one input but approximate the complex interaction of features represented by the model is come up with by marginalization over other features. This mathematical underlying provides an assurance that explanations do not occur but rather come out of a direct approach to the learned decision function. Also, these attribution scores allow to provide both local explanations, which explain single predictions, and global information, which are used to demonstrate overarching features importance trends. Consequently, explainability, based on mathematics, improves transparency, trust, and accountability in AI-based systems of decisions.

3.3. Explanation Generation Techniques

The offered methodology will combine the variety of explanation generation methods in order to offer an extensive and user-friendly insight into the model behavior on various levels of abstraction. Local explanation methods center on the individual predictions, which aim at establishing the input features that contributed most heavily to an individual output. This is especially useful when it comes to decision-critical applications because users are able to see the logic behind why this or that decision was taken concerning a specific case. Local explanations promote transparency and error analysis and also trust-calibration since they are available at the instance level, allowing a user to confirm that the reasoning of the model and domain knowledge and expectations align. Conversely, global explanation methods seek to obtain a general behavior of the model in the entire input space. Such explanations

ISSN: 3071-5628
PP: 01-11.

give an understanding of the overall pattern of decisions, prevailing trends, and acquired relationship between inputs and outputs. Global explanations need to be used to validate the model, identify bias, and ensure that the model adheres to regulations, because they enable stakeholders to understand whether the model is consistent and fair on various data subgroups. Summarization techniques including feature importance aggregation, surrogate models, rule extraction etc are applied to highlight complex model behaviour in a form that can be easily understood. In providing total picture of the decision making process, global explanations enable model debugging, and long term monitoring. Counterfactual explanations are used in complement to local and global approaches because they answer a question that an individual would rather ask, what-if. These explanations are used to explain how the changes that are minimally required in the input features can be used to change the prediction of the model to a desired solution. Counterfactuals are specifically appealing to human user, whereby they are aligned to causal reasoning and actionable decision making. As an illustration, they may provide information to the user of the specific changes in features that could change the final classification. Through the incorporation of local, global and counterfactual explanations, the methodology guarantees that explanations are technically sound in addition to being meaningful, actionable and flexible in terms of its application to different user needs and application conditions.

3.4. Evaluation Metrics for Explainability

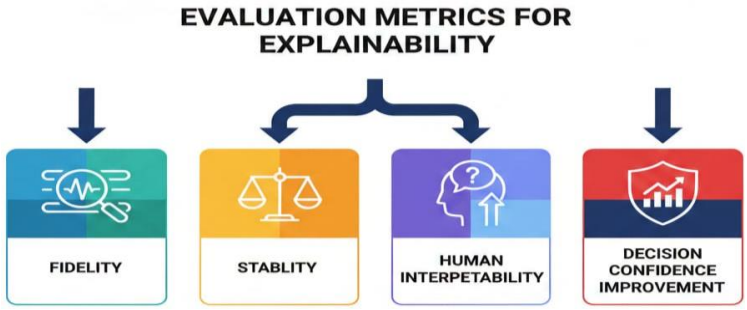

Fig 3 - Evaluation Metrics for Explainability

3.4.1. Fidelity

Fidelity is used to assess the relevance of an explanation of the predictive model that could be describing the actual behavior. High-fidelity explanation is much closer to the real decision-making process of the model and makes sure that the explanation is not false nor oversimplified to a reasonable extent. The measure is of especial interest to post-hoc methods of explanation, in which there is a risk of coming up with plausible but erroneous interpretations. The high fidelity will work to build trust by making the explanations loyal to the reasoning behind the model.

3.4.2. Stability

Stability tests how the answer to a question remains consistent when minor, significant changes are done on the input data. A good explainable measure must be that which will give similar arguments to similar inputs without sudden or conflicting arguments. The meanings are volatile and may confuse users and make them lose trust in the system. Consequently, stability is fundamental towards guaranteeing reliability and robustness particularly in real world where data changes are bound to occur.

3.4.3. Human Interpretability

Human interpretability looks at how comprehendible and interpretable to the end user the generated explanations are. The measurement values are based on the criterion of clarity, simplicity and level of correspondence with human thinking and means of knowing the domain. Technically true

ISSN: 3071-5628
PP: 01-11.

and yet hard-to-understand explanations do not serve their purpose. Human interpretability is high and this is what makes the explanations reasonably effective in terms of promoting an informed decision and user confidence.

3.4.4. Improvement in decision Confidence

The improvement in the confidence of the decision gauges the degree to which the explanation improves user confidence in an AI-assisted decision. This measure reflects whether the explanations make users feel more convinced, knowledgeable, and strict to believe model outputs. Increased confidence is especially essential in high-stakes areas, where users need to defend or take AI advice. Finally, this metric gives an insight into the practical implication of explainability on the human-AI collaboration.

4. RESULTS AND DISCUSSION

4.1. Comparative Analysis of XAI Techniques

Table 1: Comparative Analysis of XAI Techniques

Technique	Model Compatibility (%)	Local Explanation Coverage (%)	Global Explanation Coverage (%)	Interpretability Strength (%)	Computational Overhead (%)	Stability (%)
LIME	100	90	20	70	40	50
SHAP	100	85	85	80	75	80
Decision Trees	70	30	95	95	20	90
Rule Extraction	80	25	85	85	50	65

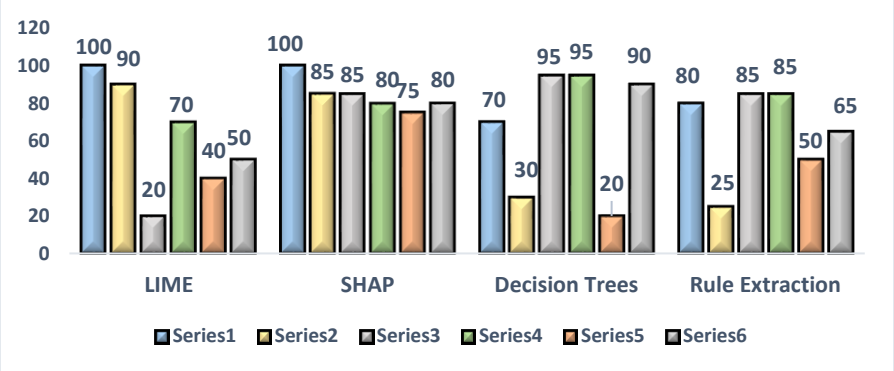

Fig 4 - Comparative Analysis of XAI Techniques

4.1.1. LIME

LIME shows complete model compatibility and, thus, can be used with a large variety of black-box models. It gives good coverage of local explanation behavior of approximating model behavior around single predictions, which contributes to instance-level interpretation. Nevertheless, its little ability to explain the world globally and relative stability implies that the explanations can be altered by minor input distortions. Although the computational overhead is rather low, one of the major limitations in critical applications is the inconsistency of explanations.

ISSN: 3071-5628
PP: 01-11.

4.1.2. SHAP

SHAP is fully compatible with black-box models and has equally good local and global coverage of explanation. It has a solid theoretical foundation that provides it with consistent and reliable feature attribution in various situations of prediction. The computational overhead is also high due to the complexity of computing Shapley values especially when using large-scale models. However, in spite of this expense, SHAP is highly stable and interpretable thus suitable in systems with high stakes.

4.1.3. Decision Trees

The interpretability of decision trees is high because decision trees are transparent and have clearly defined lines of decision-making. Their application is especially successful to elucidate the global explanations when the complete behavior of the model can be graphically depicted and comprehended. They are efficient and dependable because of the low computation cost and high consistency. There is however limited compatibility of their models and lower local explanation flexibility impairs their capacity to scale to complex and high-dimensional problems.

4.1.4. Rule Extraction

Rule extraction methods are a hybrid solution in that they synthesize rules into human-readable form based on complex models. They provide a good coverage of global explanations and a rather good level of interpretability, helping to understand the model at system level. The cost of the process of black-box behavior approximation is moderate. Stability and local explanation abilities are more or less lower, though, because of errors in approximation and simplification of rules.

4.2. Impact On Decision Transparency and Trust

Empirical evidence repeatedly shows that the inclusion of explainable artificial intelligence systems can be significantly positively concerning transparency and trust of decisions with the user. Explanations together with predictions of the model can help the users to grasp the reasoning behind the recommendations given by AI, leading to the perception of less transparency often ascribed to complex models. Such an increased transparency also means that the user is able to mentally walk through the process of decision-making, sometimes when the underlying model is complex enough to obscure, hence making the user feel in control and knowledgeably involve themselves with the system. Interestingly, empirical research demonstrates that a user trust in the decisions made with the aid of AI increases substantially when explanations are provided, in those cases, when predictive accuracy stays the same. The result of this study emphasizes the idea that the confidence in AI systems does not rely only on the performance rates but is considerably affected by the feeling of responsibility and understandability. The explanations enable users to evaluate whether model reasoning is consistent with knowledge in the domain, ethical and contextual requirements. Consequently, people become more accepting of, rationale-compelled, and responsive to AI suggestions especially in high stakes settings where choices are accompanied by greater implications. Moreover, explainable systems can enable people to work more effectively with AI because such tools allow the user to track possible errors, human biases, or incongruities in the model results. Explanations can influence critical reflection as opposed to blindly relying on automated systems by divulging influential features and patterns of decision-making. This moderated trust (also called calibrated trust) is what will help avoid an over and underuse of AI technologies. Further, transparent descriptions are associated with creating accountability within an organization since stakeholders can audit, as well as justify, automated decisions. In general, the addition of explainability supports the increased user trust and acceptance, as well as the wider legitimacy and ethical implementation of AI-driven decision systems.

ISSN: 3071-5628
PP: 01-11.

4.3. Discussion On Trade-Offs and Limitations

Though explainable artificial intelligence is much more effective in terms of enhancing transparency and trust, multiple trade-offs are bound to be introduced, which should be managed. The complexity of generating explanations, especially under post-hoc methods used on black-box models, is another major problem, due to the extra computational cost. These methods (e.g., feature attribution and surrogate modeling) typically need no fewer than repeated evaluations of the model or auxiliary optimizing processes that may raise the cost of inference in time and resources. This is particularly debilitating in real-time or large-scale systems, but efficiency and responsiveness are absolutely necessary. The other factor which is critical as a drawback is the complexity of the explanations themselves. Explanations that are very detailed can be technically precise, but may be inaccessible to all users, especially those lacking special expertise in the domain or understanding of the mechanics of machine learning. When too much information is presented it is addressed it may not be understood and even not be useful effectively in explaining things and instead, it must confuse instead of clarify. On the other hand, simplified explanations can lose fidelity that they leave an incomplete or false description of the reasoning in the model. These simplifications may give one a false impression of knowledge and even destroy trust in the event the differences between the explanations and the actual model behaviour are later found. Striking a balance between the explanation fidelity, interpretability, and usability is an unresolved research problem in XAI field. The right degree of explanation would be contextual and would change according to the level of expertise of the user, the criticality of the decisions and the field of application. Also, there is no single norm of how to evaluate the quality of explanations, and it is challenging to compare methodologies without objective comparison. These shortcomings point to the necessity of adaptive, user-friendly explainability systems that can dynamically distribute explanations to the needs of stakeholders. These trade-offs cannot be ignored, and this would address the issue of assure explainability augments, instead of harming, the effectiveness and reliability of AI-driven decision systems.

5. CONCLUSION

The paper has provided a keen systematic assessment of the occurrence, development and increased importance of Explainable Artificial Intelligence in the contemporary decision-making systems. Due to the growing impact of the artificial intelligence infiltrating the key areas of society, including the healthcare system, finance, autonomous systems, and the public policy, the need to ensure transparency, accountability, and ethical standards is now of utmost importance. Explainable AI is part of the solution to these concerns, as it helps stakeholders to comprehend, assess, and have faith in automated decisions, thus changing AI into a human-focused decision assistance framework instead of the performance-driven technology. The preliminary literature review of this study revealed the conceptual difference among interpretability and explainability and also the various types of XAI methods, which start with intrinsically interpretable model types to post-hoc methods of explanation. The suggested methodological framework revealed the systematic way explainability can be applied throughout the AI lifecycle, starting with data preprocessing and model selection and continuing with the generation of explanations as well as human-in-the-loop assessment. In addition, the comparative study of XAI methods demonstrated that even though sophisticated methods of explanation play an essential role in increasing the transparency of the decisions made and increasing their user trust, they present certain trade-offs in terms of computational cost and explanation stability, as well as, the complexity of such explanations. These outcomes and discussion emphasize the fact that explainability does not only increase transparency; it also leads to the improved robustness of system through finding the errors and identifying the bias, as well as providing the human control. The calculated trust born by XAI promotes successful collaboration between humans and AI and not the unquestioned

ISSN: 3071-5628
PP: 01-11.

dependence on computer technologies. The research also, however, states that explainability is relative in nature, and it does not have a single solution, which can be applied to all its instances, or to users in general. There is a need to focus on future research directions towards the standardization of the metrics of the evaluation of the explanations to be able to provide the objective comparison and benchmarking of the XAI methods. The domain-adaptive and user-ware explanation strategies will also have to be developed to handle various needs of stakeholders. The other promising area of relevance and usability is the integration of continuous human feedback into the processes of explanation generation. Finally, Explainable Artificial Intelligence will stand in the center of establishing responsible, transparent, and human-friendly intelligent systems that can help to balance the progress of technology and the values and ethical principles of society.

REFERENCES

[1] A. Adadi and M. Berrada, "Peeking inside the black box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[2] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.
[3] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
[4] C. Molnar, *Interpretable Machine Learning*, 2nd ed., Berlin, Germany: Leanpub, 2022.
[5] R. Guidotti *et al.*, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.
[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.
[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
[8] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Machine Learning*, vol. 102, no. 3, pp. 349–391, 2016.
[9] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proc. ACM SIGKDD*, 2016, pp. 1675–1684.
[10] D. Gunning *et al.*, "XAI—Explainable artificial intelligence," *Defense Advanced Research Projects Agency (DARPA)*, Tech. Rep., 2019.
[11] A. Holzinger *et al.*, "What do we need to build explainable AI systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.
[12] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.
[13] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*, Cambridge, MA, USA: fairmlbook.org, 2019.
[14] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.
[15] J. W. Crandall *et al.*, "Human–AI collaboration and explainable autonomy," *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 59–67, May–Jun. 2018.