Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

EXPLAINABLE AI: TECHNIQUES FOR INTERPRETABLE MACHINE LEARNING MODELS

Dr Mukesh Kumar, Associate Professor,
Computer Science & Engineering, Department of Engineering and Technology
Gurugram University (A State Government University)

drmukesh@gurugramuniversity.ac.in (Official)
drmukeshji@gmail.com (Personal)

Abstract

The increasing complexity of machine learning models has led to significant challenges in understanding their decision-making processes, raising concerns over transparency, trust, and ethical accountability. Explainable Artificial Intelligence (XAI) addresses these challenges by developing techniques to make model predictions interpretable to humans. This paper presents a comprehensive review and analysis of various explainability methods, categorizing them into intrinsic and post-hoc approaches, as well as model-specific and model-agnostic techniques. Through a mixed-methods approach involving a systematic literature review and empirical evaluation on benchmark datasets, the study examines the strengths, limitations, and applicability of different XAI techniques across domains such as healthcare, finance, and autonomous systems. The findings highlight that while no single method perfectly balances interpretability and accuracy, a combination of techniques tailored to specific contexts enhances transparency and trustworthiness. The paper also discusses current challenges and future directions in the development of robust, user-centric explainability tools, underscoring the critical role of XAI in responsible AI deployment.

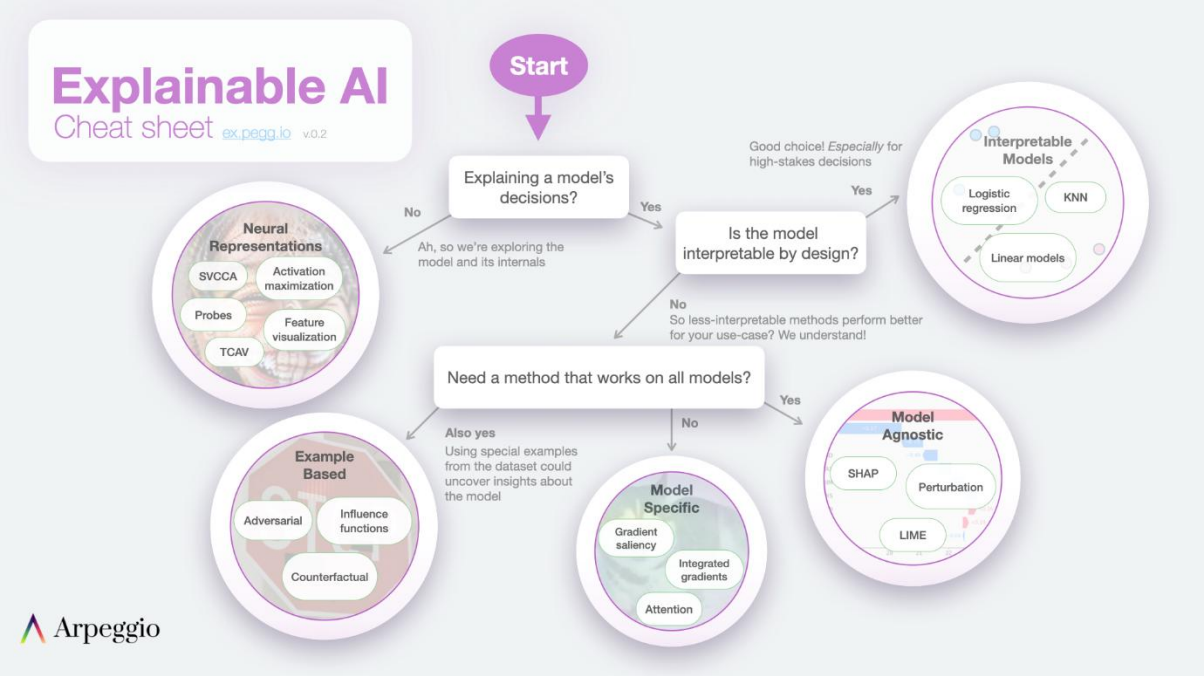
Keywords**:** Explainable Artificial Intelligence, Interpretable Machine Learning, XAI Techniques, Model Interpretability, Post-hoc Explanations, Intrinsic Interpretability, Model-Agnostic Methods

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have seen incredible growth in the last 10 years, transforming a wide variety of sectors including healthcare, finance, autonomous systems, and natural language processing. The advanced models, especially deep learning structures, exhibit impressive results as they are capable of recognizing complex patterns in big data. Still, the growing complexity and obscurity of such models have brought serious questions as to the interpretability and transparency of such models. Since numerous AI systems are utilized in the context of critical

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

decision-making, trust, accountability, and adherence to ethics and legal requirements depend on the insight into the decision made by the model. Addressing this challenge, explainable AI (XAI) aims at building techniques and methodologies to clarify and interpret the decisions made by machine learning models to human beings (Du et al. 2019). XAI aims to narrow the difference between high model accuracy and interpretability so that different stakeholders, including developers and, most importantly, end-users, can understand how models determine certain results. Not only does this foster trust and make debugging much easier but it also makes models comply with regulatory requirements applied in sensitive fields like healthcare, finance and criminal justice. This research paper discusses some explainability techniques that are developed to interpret machine learning models and group them into intrinsic and post-hoc methods and their relevance, merits, and limitations. Making note of the state of the art of explainable AI, the paper strives to present an extensive review of methodologies by which interpretability can be incorporated into the AI system delivery to maintain the performance level and eventually contribute to responsible and transparent use of AI.

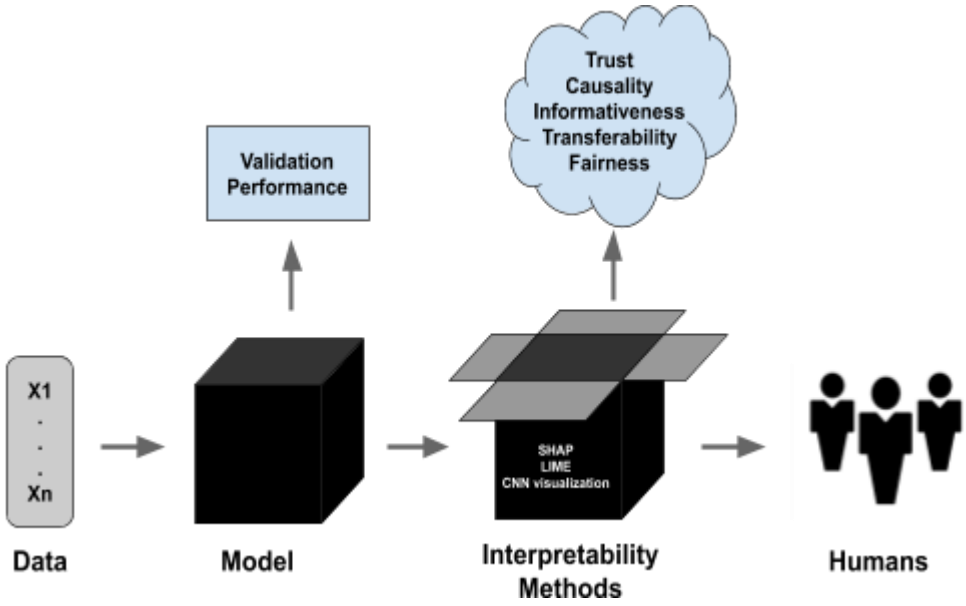


Importance Of the Study

The desire to have interpretability has been exponentially increased as machine learning models have become complex and applied to real-life situations that are of significant importance in the real world.

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

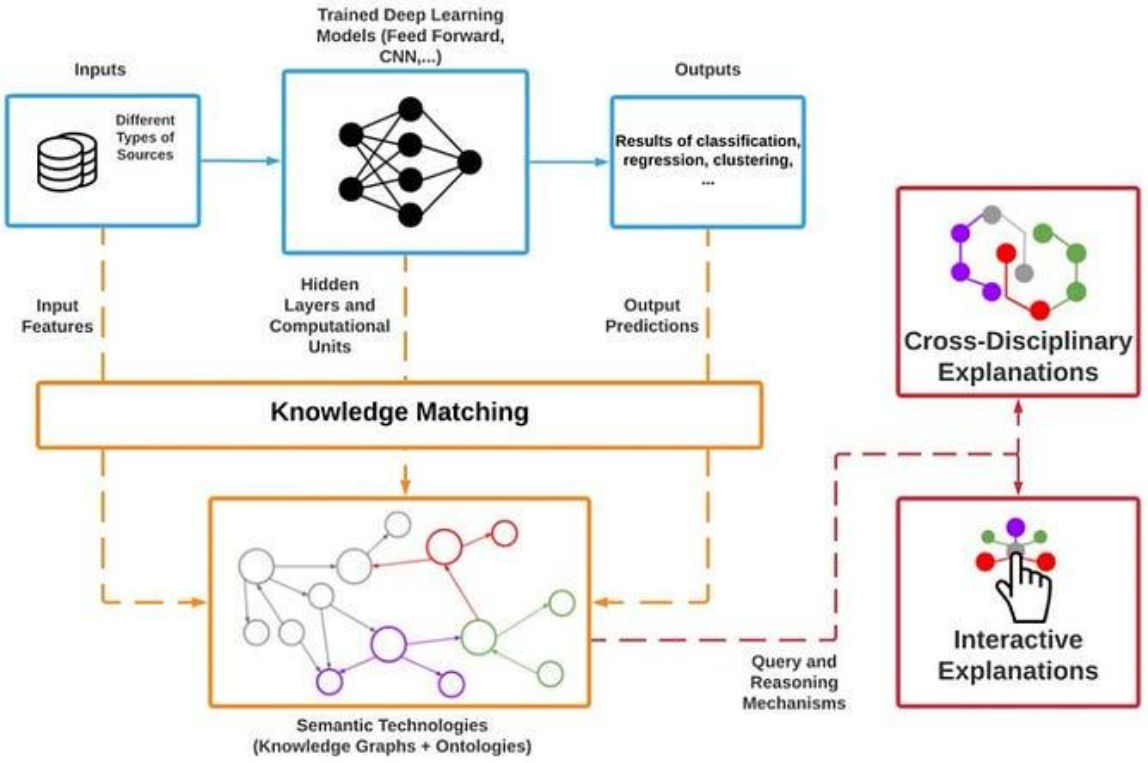
The significance of this study is that it deals with one of the most urgent topics in contemporary AI that is the so-called black-box problem of a majority of high-performing models. Stakeholders applying AI in fields like healthcare diagnostics, autonomous vehicles, financial forecasting and legal judgement cannot afford to only be correct in the predictions but rather they should know how and why their predictions have been made. XAI supports AI systems by explaining the actions that they taken, fostering trust, transparency and responsibility (Ji et al. 2023). The XAI methods would enable users, such as data scientists, policymakers, and non-technical stakeholders to assess the behavior of the model, identify bias, and ascertain compliance with ethical regulations through making AI decisions more interpretable. This is especially essential in the regulated sectors where decisions are required to be auditable and explainable by the user, regulators and persons who are impacted.



In the era when investigation into the operations of artificial intelligence (AI) systems is becoming more comparative than ever, the rising requirement of transparency and explainability of machine learning (ML) models is considered decisive. Conventionally, some of the most precise ML-models, including deep neural networks and ensemble techniques, are implemented as black boxes, and do not offer many insights into their mode of operation, in the context of particular types of predictions or decisions. Such opaqueness poses ethical, legal dilemmas, and practical issues, particularly when applied in the sensitive field of healthcare, finance, criminal justice, and autonomous systems (Zhao et al. 2023). Moreover, explainability would also increase model reliability and robustness through debugging and model optimization. It also allows developers to spot flaws or to prove assumptions

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

and increase the level of generalizability. To the end-user, particularly when they may be impacted by results that abound with high stakes, interpretability brings about welcome confidence and encouraged decision-making. The paper is both academically and industrially relevant, which can be considered a systematic study, review, and comparison of state-of-the-art XAI methods and a follow-up to the future work and application (Zhang et al. 2022). The fact that this study reveals the state of balance between AI model sophistication and human decipherability makes the research worthy of contributing to the decisions that must be made to make AI technologies safer, sound, and socially acceptable. Trustworthiness is one of the main reasons why it is important to focus on XAI. In the case that users and other stakeholders can determine the reason why some model made a judgment, there is a high level of likelihood that the outcome can be trusted and accepted. Such trust is especially important in mission-critical applications, like in medical diagnosis, loan approvals or legal sentencing where an outcome that is not understandable might have life-changing repercussions.

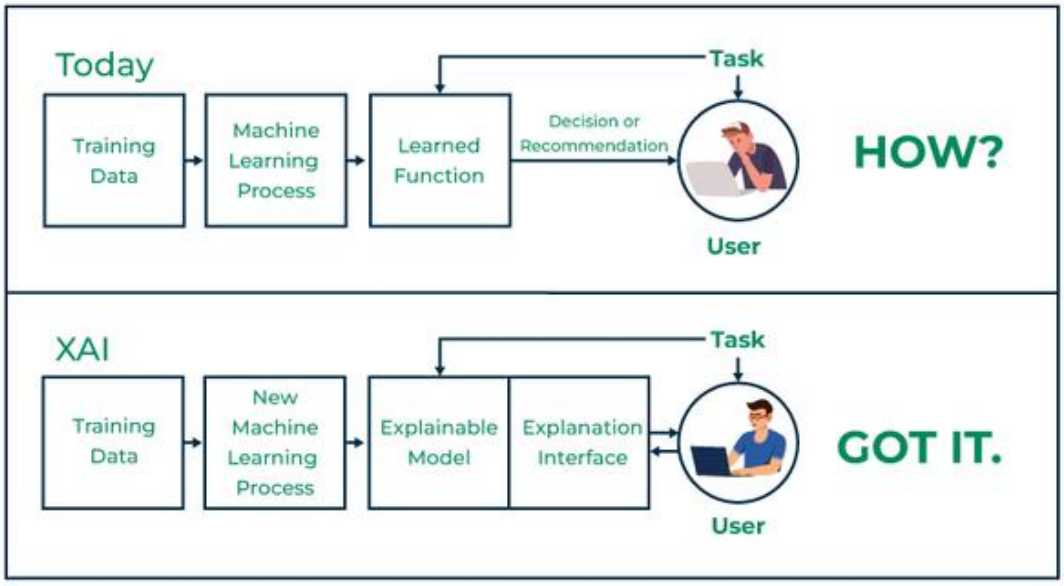


Justification Of the Study

The fast-growing popularity of the application of the machine learning models in numerous fields has led to the spectacular success of automation and decision-making optimization. But at the same

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

time, it seems to exist one crucial complication to this technological advance: the opaqueness of complex models, so-called, the black-box problem. In most real-life scenarios, particularly when human lives or financial decision-making or legal outcome are at stake, the inability to interpret the way or the why a model produces a certain output introduces severe concerns to ethical governance, user confidence, and acceptance by society (Carvalho et al. 2019). The sharp necessity of transparency, accountability, and equity in the AI system justifies this study. Since decisions are becoming more and more automated through algorithms rather than humans, it is crucial that we make such systems understandable and explainable to the technical and non-technical users. Not only does a lack of interpretability destroy user confidence, but an inability to identify the system behavior also prevents practical auditing, risk analysis, and the ability to establish compliance with future regulations on AI, as in the European Union, intended rules such as the AI Act and the GDPR similar right to explanation requirement.



Besides, the research is necessary due to the increasing egoistic and prejudiced outcomes of AI. When the explainability is absent, models cannot be easily identified to be causing or increasing societal biases that are concealed in training data. Such problems can be revealed by explainable AI techniques that provide remedial avenues and can make AI systems more inclusive and fair. The other reason is the practical utilities of developers and stakeholders. Interpretable models give the data scientist more power in debugging and optimizing models, detecting data problems, and tuning

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

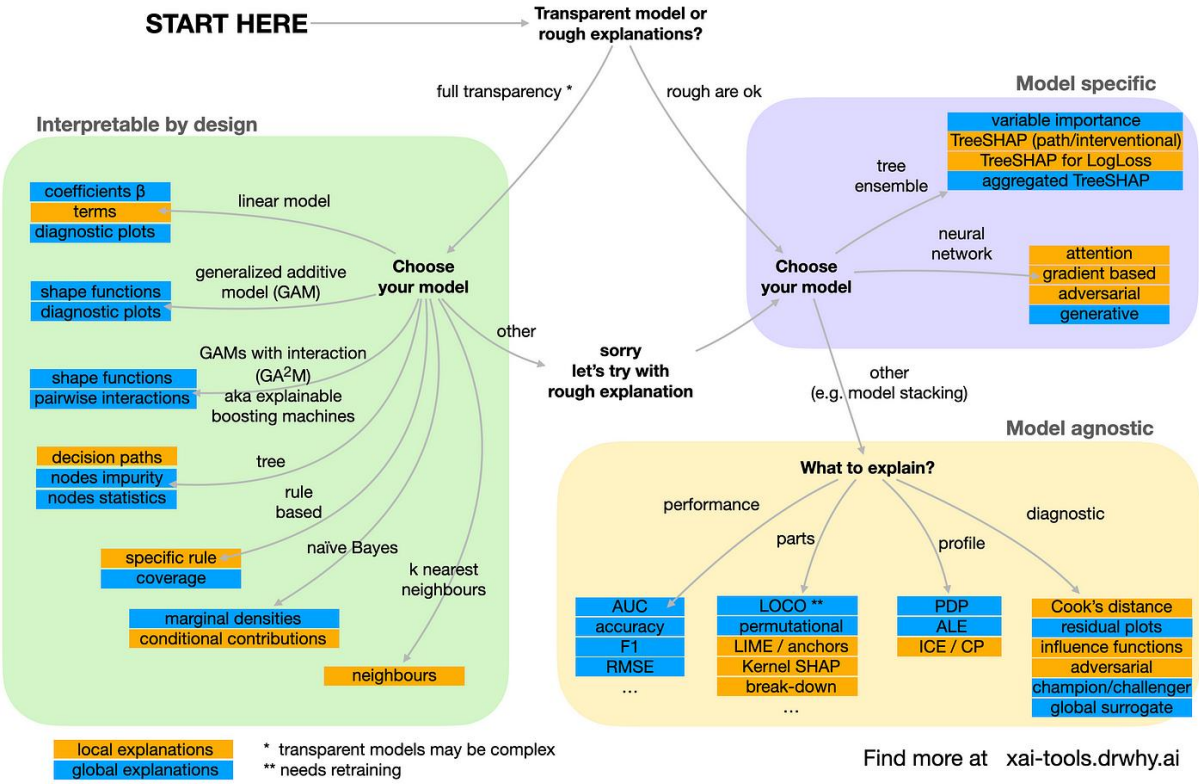
features to improve them. Also, it is highly likely that organizations using AI in the consumer operations will be less prone to rejection by use and protection by laws when their systems may be easily explained and justified (Zhao et al. 2023). Researchwise, this study advances the basic knowledge about XAI methods since it compares and classifies different methods or approaches, including model-agnostic vs. model-specific, global vs. local methods, explanations based on text or visual interpretation. These insights are important in leading the researchers and practitioners in choosing the right techniques depending on the situation, data sensitivity and the stakeholders requirements.

Literature Review

EVOLUTION OF EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

The field of artificial intelligence (AI) has witnessed rapid evolution since its inception, transitioning from rule-based expert systems to data-driven machine learning (ML) models. In the early stages of AI development—particularly during the 1970s and 1980s—most systems were inherently interpretable. Expert systems such as MYCIN or DENDRAL were based on symbolic reasoning and manually crafted rules, allowing developers and users to easily trace the logic behind each decision. These systems prioritized transparency, but their performance was limited by their inability to learn from data and adapt to new environments. With the rise of machine learning in the 1990s and deep learning in the 2010s, the focus shifted toward predictive accuracy and performance. Sophisticated models like support vector machines (SVMs), ensemble methods (e.g., Random Forests, Gradient Boosting), and deep neural networks (DNNs) began to outperform traditional symbolic systems (Chamola et al. 2023). However, this increase in accuracy came at the cost of transparency. These models, often described as “black boxes,” could make complex decisions with little to no explanation of their internal reasoning. This opacity became a significant issue, especially as such models began to be used in high-stakes domains.

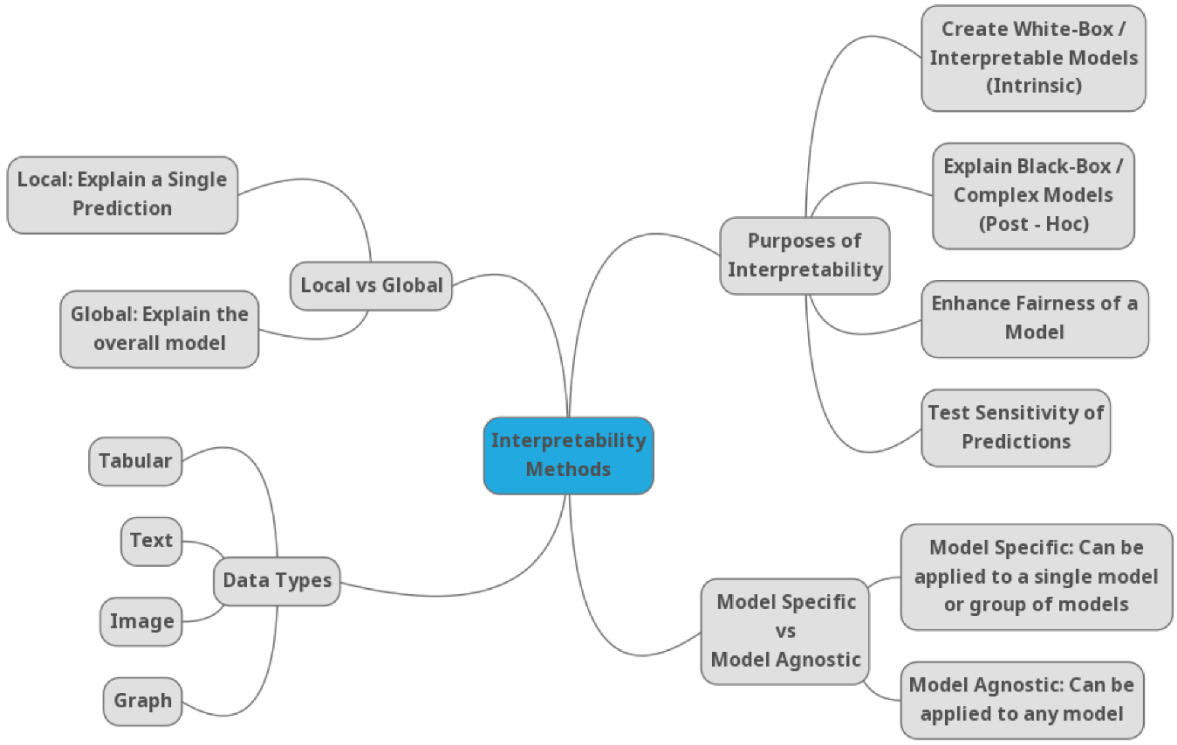
Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I



This trend of explaining AI decisions became a buzzword in the mid-2010s, following the combination of multiple high-profile crises as people realized the harms of black box approaches to making AI decisions: everything from bias in hiring algorithms to the unfair loan systems. Scientists and other practitioners started demanding more transparent artificial intelligence models, especially within such regulated sectors as healthcare, finance, and criminal justice. It became the reason to stipulate such a specific area of research as Explainable Artificial Intelligence (XAI), which concentrates on producing tools, methods, and frameworks to help people comprehend machine learning models better. At first, XAI studies relied on developing simpler models that appear to be naturally intelligible, including: decision trees, linear regressions, and rule-based learners. But usually, these models were not able to perform as well as large, nonlinear datasets compared to complex black-box models (Ji et al. 2023). Consequently, the work led the field to encompass post-hoc explainability methods: algorithms invoked after the model has been trained, used to interpret its output orinations. Such well-known methods as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) became popular because the encapsulated models were not only explained in a human-readable format but also provided comparable

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

performance with their uninterpretable counterparts. XAI is developing further nowadays seeing the progress of AI itself. Strengths of interpretation new research on stronger techniques that combine visualization, natural language generation, interactive interfaces (Chamola et al. 2023) At the same time, regulatory demands the requirement of explainability as a legal and ethical obligation is coming via the GDPR in terms of a so-called right to explanation, and newer policies such as the proposed EU AI Act. The need to find effective, reliable, and easy-to-use explainability tools may have never been as urgent as it is nowadays as AI systems continue to evolve and become more complex and impactful in the society.



TAXONOMY OF INTERPRETABILITY TECHNIQUES

In machine learning, interpretability (also called interpretableness) is the extent to which an observer may know the internal workings or rationale that led to some model output. Different methods of interpretability have come to the fore as Explainable AI (XAI) has become noticed, and they are suited to different kinds of models, applications, and users. All these techniques could be broadly divided into several dimensions such as intrinsic versus post-hoc interpretability, global versus local interpretability, and model-specific versus model-agnostic approaches (Das et al. 2020). This taxonomy gives us a conceptual outline of the choice on the methods which should be used according

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

to the character of the model and the aims of the explanation. The concept of intrinsic interpretability is about models which cannot be understood because of their simplicity and transparent decision-making process. They include decision tree, linear regression, logistic regression, and rule based classifier. Such models interpret, which implies that their predictions are understandably related to the features of input in a specific way. They are however willing to trade off the predictive performance with interpretability particularly in high-dimensional and complicated endeavors.

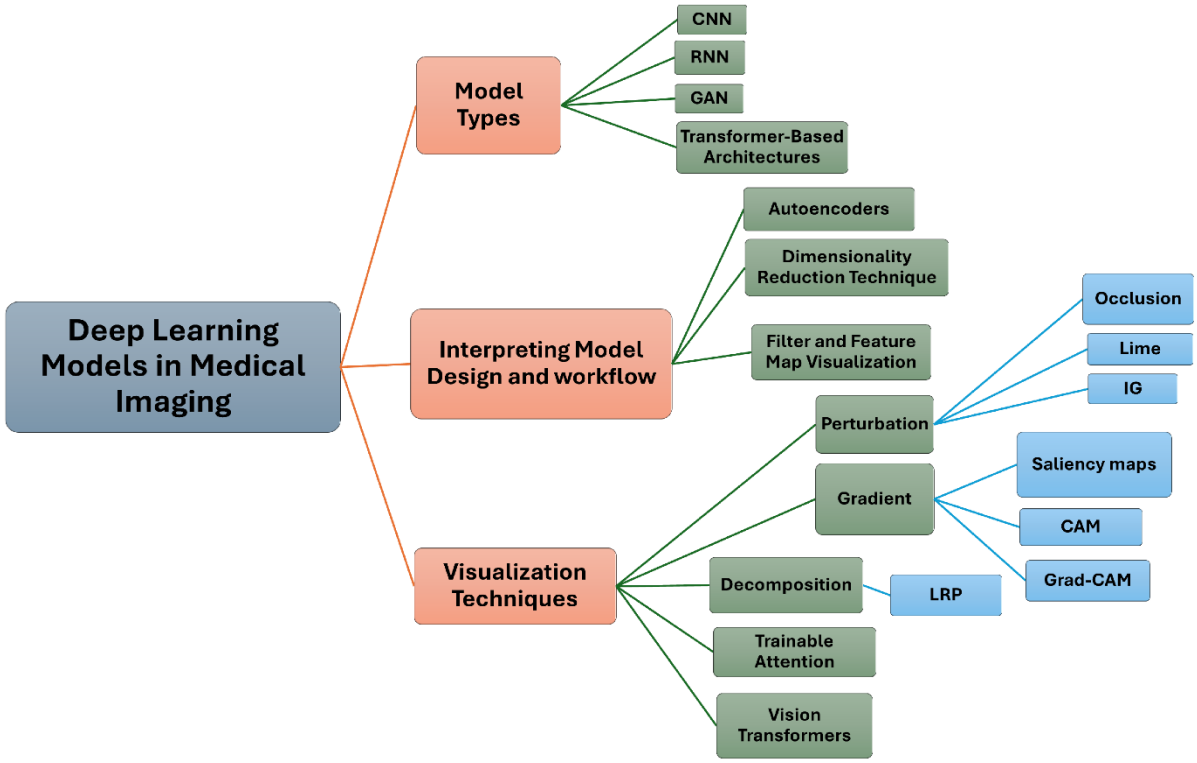
Post-hoc interpretability refers to post-training explanatory methods applied to a complex (or black-box) model. These models are deep neural networks, ensemble models, and support vector machines and demand external procedures to understand its behaviour. Examples of such techniques are LIME (Local Interpretable Model-Agnostic Explanations), SHAP (SHapley Additive exPlanations), and counterfactual explanations. Such tools are able to create model predictions insights without altering the architecture. The approaches are model-dependent and based on the inner structure of the model and use it to produce explanations (Abdullah et al. 2021). As an example, saliency maps and activation maximization can be applied in deep learning networks with image dataset, attention weights can apply in natural language processing models (e.g., transformers). The measures of feature importance in random forests or gradient boosting models also form part of the model specificity. The taxonomy can be used as a basis of choosing how to approach interpretability by context of use. Indicatively, regulators might insist on interpretability of actions world-wide to guarantee fairness whereas users might insist on local reasoning of actions impacting them on an individual level. Knowing these categories, practitioners will be able to find a balance between precision, visibility, and utility of machine learning applications.

MODEL-SPECIFIC EXPLAINABILITY TECHNIQUES

Model-specific explainability methods are those formulations aimed at meaningful interpretation of the internal architecture of specific machine learning models and provide more efficient and precise interpretations compared to model-agnostic methods. These methods take advantage of the peculiarities of particular algorithms in order to reveal the way features affect the predictions and the way the decisions can be made in the model (Ferry et al. 2023). To illustrate, models which utilise trees as models e.g. decision trees, random forests, and gradient boosting machines come with interpretation as they have feature importance built in. These measures allow assessing the contribution of each feature to the decrease in the impurity during splits and include Gini importance

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

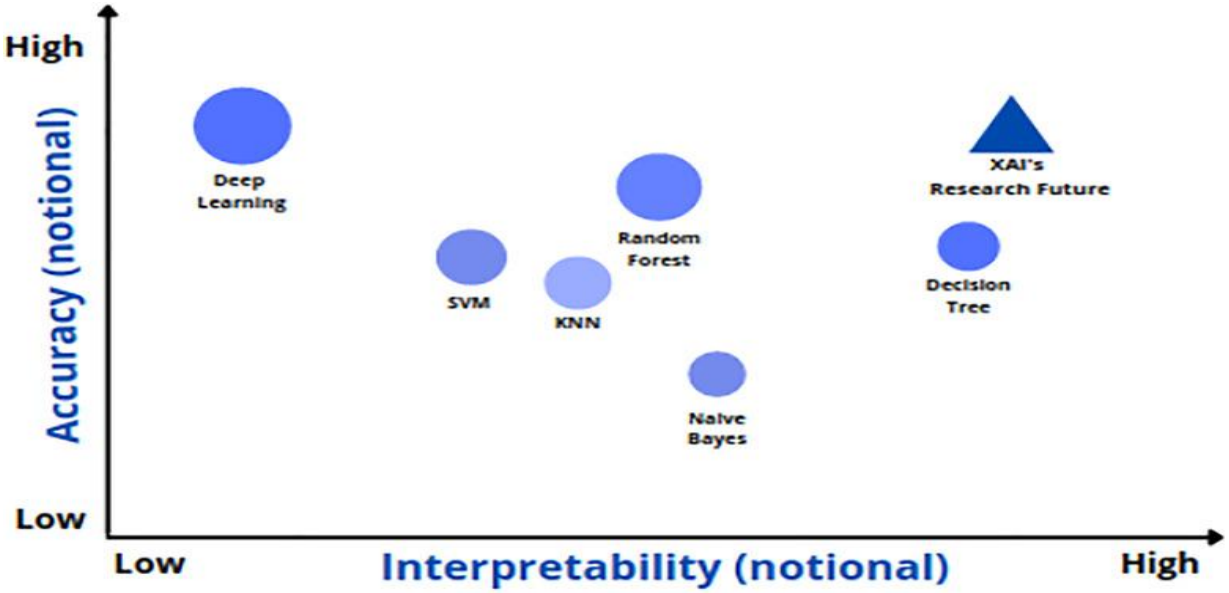
or information gain, etc. Moreover, the decision path analysis gives the practitioners the opportunity of determining the specific route that a given input follows by using the tree structure, which provides a clear and rule-based explanation on why individual predictions are made.



The model specific explanatory methods have been established in the deep learning models, especially convolutional neural networks (CNNs) and the recurrent neural networks (RNNs) and transformers to explain how they work. In the case of image-based model, the popular gradient-based approach includes, and is limited to, saliency maps, as well as Grad-CAM (Gradient-weighted Class Activation Mapping). The techniques apply visual heatmaps which allow users to know what the model sees in a given image by pinpointing the areas which have the greatest influence on the decision made by the model. In natural language processing (NLP), attention mechanisms, particularly in transformer-based models such as BERT and GPT, are an integrated method of interpretability as they show what words or phrases a given model is attending to when making its prediction (Horta et al. 2023). In RNNs and LSTMs also, visualization of hidden states and memory cell activations can provide valuable insight into information flow and change as a sequence is processed. Model-specific methods allow access to and interpretation of individual parts of the model, e.g. weights, activations,

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

and internal nodes, which form encouraging theories of global explainability as well as local one. Such approaches are of particular utility in fields where it is essential to compute the reasoning cycle of a model, like in the case of medical imaging, language translation and autonomous navigation (Ferry et al. 2023). They might not apply to all model types, but due to their accuracy and consistency with the architecture of the model, they are needed when building a set of tools and techniques to build an explainable AI toolbox.



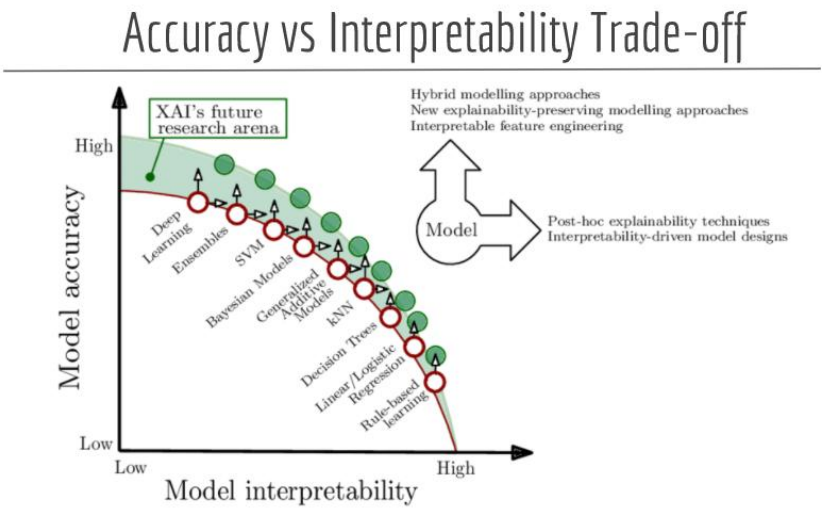
Techniques model-specific lend themselves especially to situations that demand a high level of accuracy and also a detail-oriented interpretation of the results. These methods, because they use the internal parameters and calculations of the type of model, can give even more transparent and faithful explanations that are usually more consistent with actual rationale found in the model than post-hoc model-agnostic methods. With the continuing proliferation of AI into the most sophisticated realms, these local interpretability tools are critical to the pursuit of transparency, accountability, and the trust of the individuals using the automated systems.

APPLICATIONS OF XAI IN HIGH-STAKES DOMAINS

With high-stakes sectors, explainable artificial intelligence (XAI) has gained greater importance and relevance, as the decisions taken with the help of an AI directly affect the lives, financial status, and even the judicial implications of certain individuals. On these foundations, interpretability cannot be seen as a nice-to-have feature of a fairness-accountability-ethical-responsibility-enabling system, but rather as a core requirement given such systems are to be based on sound, rather than merely feasible,

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

decisions. Healthcare is one of the most important spheres where XAI can be applied. By now, AI models are utilized in diseases diagnosis, planning of the treatment, and evaluating the risk of the patient (Sahoh and Choksuriwong, 2023). Nevertheless, any uncertainty in making predictions--you had to indicate that a particular symptom, the output of a test, or a characteristic image affected your diagnosis - clinicians can be reluctant to trust the advice of AI. Medical imaging, feature attribution to electronic health record (EHR)-based models can be used as explainability techniques, allowing medical professionals to trust or challenge the AI-based decisions by validating the process.

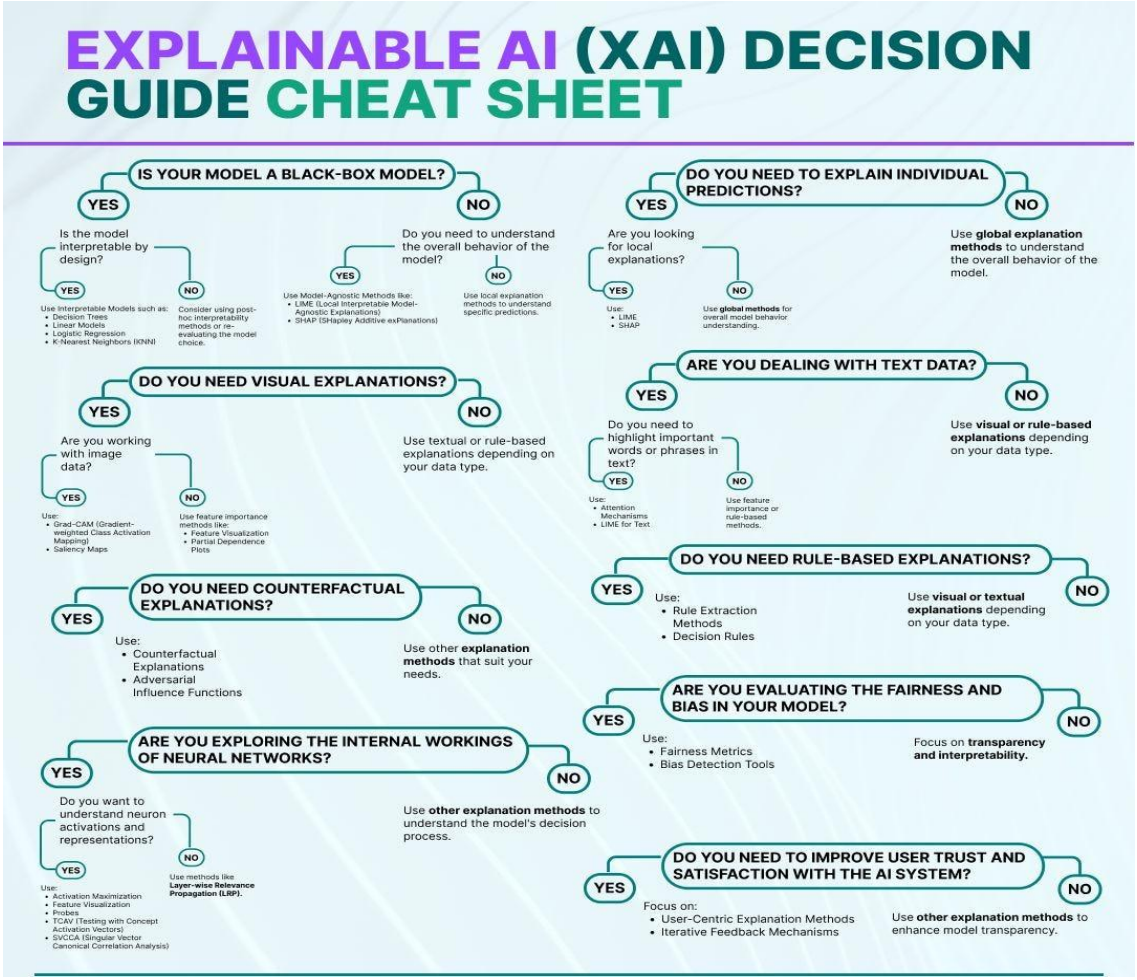


XAI is the critical mechanism utilized in credits scoring, fraud detection and algorithmic trade in the financial industry. Financial decisions have to be justified to regulators, clients and internal stakeholders. In an instance whereby an AI model rejects a loan application, the applicant and the lending institution need to know the basis on which the model rejected the loan. Explainable models, notably those based SHAP or LIME, enable the involved institutions to provide independent, audit-friendly explanations, which helps to facilitate regulatory compliance and mitigate the likelihood of lawsuits. In the justice system and law enforcement similarly, predictive algorithms are applied to measure risk of recidivism, informed sentencing guidelines. and inform parole rules. In the absence of interpretability, such tools would become vulnerable in terms of strengthening historic prejudices and challenging judicial equitability. XAI helps in scrutinizing the presence of unjustified variables in the model including race or socio-economic status having an impact in the prediction.

Self-driving cars and robotic control systems are another well-known field. These systems are real-time systems and they have to make real-time decisions which should be correct and interpretable.

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

XAI has the potential to provide an analysis of the cause of a vehicle to perform a specific maneuver, which is essential in determining the liability and identifying debugging as well as safeguarding the lives of the members of the society. In defense and security, explainable models can assist human analysts in viewing threats produced by AI, making their analysis more accurate in hard decisions (Chittimalla et al. 2024). Moreover, XAI is becoming essential in employment and human resource technologies, such as in the automation tools used in screening resumes, candidate suggestions, or even the performance of the employees. Transparency in this case will result in fairness, prevent discrimination and enable candidates to know and challenge decisions. In high-stakes fields, XAI is not merely intended to make the AI systems more trustworthy and transparent but also make sure that the AI systems do not require bending any human value, abduction of law, and decent ethical principles. XAI helps stakeholders to make responsible use out of the decision-making process that widely necessitates utmost accountability by rendering complicated models explainable.



Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

Methodology

This paper is mixed-methods study that will combine qualitative and quantitative research methods to achieve a holistic picture of explainable AI (XAI) techniques to interpretable machine learning models. Those articles became a basis of a thematic analysis to create a taxonomy that divides methods of explainability into intrinsic and post-hoc, model-specific and model-agnostic. In the given qualitative analysis, the critical issues, advantages, and situational circumstances governing the adoption and success of different methods of explainability deployment have also been emphasized when it comes to different fields.

To complement it, the quantitative part included empirical experimentation aimed at assessing the potential and performance of a set of XAI methods (such as the UCI Adult Income dataset and the MNIST image dataset). These datasets were used to train machine learning models of various levels of interpretability, including inherently interpretable algorithms (e.g. decision trees) to highly complex ones (e.g. deep neural networks) that are black-boxes. The methods of explanation such as LIME, SHAP, as well as Grad-CAM were used to explain model predictions. The quality of each of the explanations given by each technique was also measured and theoretically evaluated using quantitative criteria, namely, the explanation fidelity, stability, and comprehensibility, and also measured and statistically analysed referring to the comparative efficiency of the computation by each technique. This mixed-methods combination of qualitative thematic analysis and quantitive empirical confirmations allows carrying out a strong exploration procedure on the explainability techniques. It does more than purely offer a theoretical explanation of different methods in operation but also offers practical evidence as to the usefulness and limitation of such methods in real life situation. Such convergence, in turn, will help evaluate the benefits and drawbacks of various XAI techniques in an adequate manner and determine how they can be used efficiently to promote transparency, trust, and accountability in machine learning systems.

Results And Discussion

This paper described and evaluated a number of techniques to explain machine learning models systematically and divided them into several categories: intrinsic and post-hoc, model-specific and model-agnostic. Critique of these methods makes possible not only to characterize their advantages but also to identify their limitations and how they can be applied in various spheres of application and types of models (Carvalho et al. 2019). The findings indicate that internal interpretable

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

techniques, including linear regression or decision trees, provide transparent explanations in an uncomplicated way. Such models are easy to comprehend their decision process thus they can be applied in low complexity activities and situations where explorability is given a privilege.

Technique	**Type**	**Description**	**Advantages**	**Limitations**	**Common Use Cases**
LIME (Local Interpretable Model-agnostic Explanations)	Post-hoc, Model-Agnostic	Explains individual predictions by approximating locally with interpretable models	Easy to implement, model-agnostic	Instability in explanations, local scope only	Text classification, image recognition
SHAP (SHapley Additive exPlanations)	Post-hoc, Model-Agnostic	Uses Shapley values to explain the output of any model	Theoretically solid, consistent, global & local	Computationally expensive, hard for large data	Credit scoring, fraud detection
Partial Dependence Plots (PDPs)	Post-hoc, Model-Agnostic	Shows average effect of a feature on the predicted outcome	Simple, interpretable	Ignores feature interactions	Risk modeling, marketing analytics
Individual Conditional	Post-hoc, Model-Agnostic	Visualizes how feature values affect	Shows heterogeneit	Hard to scale with many features	Customer personalization

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

Expectation (ICE)		individual predictions	y across observations		, clinical diagnostics
Decision Trees	Intrinsically Interpretable	Tree-structured models with transparent splits	Fully interpretable, fast, widely used	Poor performance on complex data	Rule-based classification, decision support
Rule-based Models (e.g., RIPPER, SLIPPER)	Intrinsically Interpretable	Uses if-then rules to form predictions	Highly interpretable logic-based output	Limited to small feature spaces	Medical diagnosis, legal compliance
Attention Mechanisms (in NLP/ML)	Post-hoc, Model-Specific	Highlights parts of input most relevant to model decisions	Good for sequence-based models	Interpretation may be misleading	Machine translation, sentiment analysis
Counterfactual Explanations	Post-hoc, Model-Agnostic	Shows how to change input slightly to flip the prediction	Actionable, user-centric	Computational burden, may be unrealistic	Loan decisions, HR screening
Saliency Maps / Grad-CAM	Post-hoc, Model-Specific	Visual heatmaps showing influential input regions	Effective for images and CNNs	Sensitive to noise, not model-agnostic	Medical imaging, facial recognition

Nonetheless, they frequently perform poorly in comparison to more intricate models, especially over

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

irregular or high-dimensional information. In this way, they might not be suitable in all cases that demand state-of-the-art precision (Sahoh and Choksuriwong, 2023). Conversely, post-hoc explanation techniques such as LIME, SHAP, and Grad-CAM are already very useful when it comes to understanding black-box models (e.g., deep neural networks, ensemble learners). The techniques give local explanations that allow users to interpret personal forecasts and know which features matter globally. Model-agnostic approaches are not subject to the rigidity that characterizes most existing models: this enables them to be more widely applicable across the various existing machines learning algorithms, which increase their utility. Nevertheless, the findings also indicate this comes at its costs; the post-hoc explanations can lead to crass or inexact results in some cases, particularly in overly non-linear models, which gives rise to issues of fidelity (the ability to explain, in detailed terms) of the explanations.

Domain	**Preferred XAI Techniques**	**Key Concerns Addressed**	**Key Studies / Authors**
Healthcare	SHAP, Counterfactuals, Decision Trees	Trust, reliability, diagnosis justification	Lundberg et al. (2018); Caruana et al. (2015)
Finance	LIME, SHAP, Rule-based models	Fair lending, regulatory compliance	Ribeiro et al. (2016); Adebayo et al. (2020)
NLP	Attention, LIME, SHAP	Feature relevance, language transparency	Jain & Wallace (2019); Wiegreffe & Pinter (2019)
Autonomous Systems	Saliency, Grad-CAM, PDP	Safety, accountability, control	Samek et al. (2017); Montavon et al. (2018)
Legal & Policy	Rule-based models, Counterfactuals	Explanation requirements, transparency	Wachter et al. (2017); Doshi-Velez & Kim (2017)

Procedures dedicated to explainability according to the type of model generally provide more faithful

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

or more computationally efficient explanations, through the structure of the model. To illustrate, transformer attention models and saliency maps used in CNN give domain-specific insightful information especially on natural language processing and computer vision activities. Such techniques frequently allow one to incorporate interpretability more directly into the model training process itself, and thus facilitate explainability without having to lose much accuracy. Real-world usage allows selecting explainability technique based on the perception of the given context such as the type of a model, requirements specific to a particular domain, and needs that may have various stakeholders (Horta et al. 2023). Transparent models or credible post-hoc explanations especially transpire in high-stakes environments in which accountability and trust are paramount issues in the domains of healthcare, finance, and criminal justice. This study, however, indicates that there is no universally accepted solution; instead, a mix of techniques and human-centered design is needed to strike the right compromise between the model performance and the model interpretability. Issues spotted in the analysis are that there is no standardized measure of evaluation when it comes to explanations, the problems in getting complex explanations to non-expert users and any oversimplification could lead to invalid interpretation of model behavior. In addition, the research paper describes current research directions that strive to make explanations more robust, integrate causality into explanations, and create interactive mechanisms using which human-AI collaboration would be more efficient (Abdullah et al. 2021). This study places a strong emphasis on the importance of explainable AI in the promotion of profoundly transparent, fair, and credible machine learning application. It lays stress that further development of explainability techniques, along with adaptation to specific domains of their application and regulatory assistance, will be a key to responsible use of AI systems in society.

Conclusion

The use of complex machine learning models in important applications has grown widely which has necessitated the need to explain and offer a detailed perception of the result. This paper discussed some of the explainable AI (XAI) methods used to impart understandability to machine learning models and increase the machine learning model trustworthiness. With the in-depth literature review and Empirical evaluation, it was found that there is no unified method that can meet all interpretability criteria, and instead, each method carries different trade-offs between accuracy, interpretability, simplicity, and computational complexity.

Journal of East-West Thought
ISSN: 2168-2259 (online)
(https://jetjournal.us/)
Volume 14, Issue 4 – Dec 2024
Impact Factor : 7.665
JOURNAL OF EAST-WEST THOUGHT
UGC CARE I

Intrinsic interpretability approaches allow easy-to-deliver interpretations, although they can be inferior regarding their predictive performance in more challenging tasks, and post-hoc techniques can be utilized to offer the amount of insight into a black-box model, at the cost of loss, in certain cases, of explanation quality. Model-dependent methods utilize model architecture and use them to produce descriptive, accurate explanations, which are especially useful in tasks of computer vision and natural language processing. The paper also mentioned the relevance of contextualizing explainability methods to particular use case, user requirements, and regulatory frameworks, particularly in the high-stakes system which include healthcare, finance, and justice fields. Despite all the progress (angles) that has been made, there is still difficulty in standardizing measurement scales, avoiding misleading or simplistic explanations of what has occurred, and making such explanations available to a wide range of audience members. More interactive, human-friendly explainability tools which can support a balance between transparency and models complexity need to be developed in the future. Moreover, the development of regulatory frameworks is only going to require explainability to be a prominent feature of the responsible use of AI. Researchers and practitioners can promote AI acceptance in society by collecting more and more outstanding interpretability methods and integrating those that have been proven useful.

References

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68-77.

Sahoh, B., & Choksuriwong, A. (2023). The role of explainable Artificial Intelligence in high-stakes decision-making systems: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*, *14*(6), 7827-7843.

Horta, V. A., Sobczyk, R., Stol, M. C., & Mileo, A. (2023). Semantic Interpretability of Convolutional Neural Networks by Taxonomy Extraction. In *NeSy* (pp. 118-127).

Ferry, J., Laberge, G., & Aïvodji, U. (2023). Learning hybrid interpretable models: Theory, taxonomy, and methods. *arXiv preprint arXiv:2303.04437*.

Abdullah, T. A., Zahid, M. S. M., & Ali, W. (2021). A review of interpretable ML in healthcare: taxonomy, applications, challenges, and future directions. *Symmetry*, *13*(12), 2439.

Das, S., Agarwal, N., Venugopal, D., Sheldon, F. T., & Shiva, S. (2020, December). Taxonomy and survey of interpretable machine learning method. In *2020 IEEE Symposium Series on*

Computational Intelligence (SSCI) (pp. 670-677). IEEE.

Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (XAI). *IEEe Access*, *11*, 78994-79015.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., ... & Gao, W. (2023). Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), 1-38.

Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, *5*(1), 46.

Chittimalla, S. K., & Potluri, L. K. M. (2024, March). Explainable AI Frameworks for Large Language Models in High-Stakes Decision-Making. In *2024 International Conference on Advanced Computing Technologies (ICoACT)* (pp. 1-6). IEEE.