Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 6, 2025

THE ROLE OF EXPLAINABLE AI IN MACHINE LEARNING MODEL INTERPRETABILITY

Muhammad Ali Khan [1], Farman Ali [2], Khadija Tahira[3], Sarah Ilyas[4], Muhammad Ahmad [5]

Muhammad Hasham Haider [6]

[1]Lecturer, School of Information Technology, Minhaj University Lahore, Lahore, Pakistan
[2]Lecturer, School of Information Technology, Minhaj University Lahore, Lahore Pakistan
[3]Senior Lecturer, School of Information Technology, Minhaj University Lahore, Lahore Pakistan
[4]Lecturer, School of Information Technology, Minhaj University Lahore, Lahore, Pakistan
[5] Lecturer, School of Information Technology, Minhaj University Lahore, Lahore. Pakistan
[6]Lecturer, School of Information Technology, Minhaj University Lahore, Lahore. Pakistan
Alikhan.cs@mul.edu.pk[1] farmanali.cs@mul.edu.pk,[2]
Khadija.tahira927@gmail.com[3],sarahillyas.cs@mul.edu.pk[4],muhammadahmad.it@mul.edu.pk[5],
hashamhaider.cs@mul.edu.pk[6]

DOI: https://doi.org/

Keywords
Explainable AI , Machine Learning Interpretability, Model Transparency, Feature Importance, Trust in AI, Black-Box Models, Rule-Based Explanations, Decision-Making.

Article History
Received on 08 May 2025
Accepted on 08 June 2025
Published on 20 June 2025

Corresponding Author: *
Muhammad Ali Khan

Abstract

This work investigates the contribution of Explainable AI to the interpretability of ML models by analyzing several methods of enabling model interpretability and how this benefit stakeholders through increased trust and usability. Explainable Artificial Intelligence has become an important part of the research area in machine learning to cope with the diamond's black box of verbose models. ML applications increasingly target sensitive sectors like healthcare, finance and law enforcement. It is making transparency and interpretability critical for building trust, enhancing decision-making and meeting regulatory requirements. This study is carried out using the mixed method, adopting the systematic literature review method alongside an empirical analysis of explainability techniques. A case study is performed in a real-world application, involving user perceptions and model performance trade-offs when using XAI methods. The discoveries clarify that while explainable artificial intelligence has procedures expanded the interpretability of a model. There was commonly a compromise between precision and explicability. This work highlights that the choice of explainable artificial Intelligence has method driven by the needs of the use case and goals of stakeholders. The task-specific efforts in developing scalable, such as, consistent, explanatory and real-time applicable. Explainable Artificial Intelligence has techniques are essential to promoting even wider integration of XAI methodology in ML-driven decision-making systems.

INTRODUCTION

1.Introduction:

The rapid advancements in machine learning decision-making is becoming increasingly efficient in various fields such as healthcare, finance, and autonomous systems. However, in many cases ML models are complex, like DL ensembles, where the model almost acts like a "black box," making it hard for researchers, practitioners, and the public to know how the models generated individual predictions (Linardatos et al., 2020). Such opacity questions the transparency, trust, responsibility, and fairness of AI leading to outcomes. This has given rise to Explainable Artificial Intelligence which is a critical field focused on making ML models more interpretable that help to understand, trust and effectively use AI systems (Kamath and Liu, 2021). AI model in the medical diagnosis domain determines that an individual patient has a particular disease. It makes comprehensible arguments about the decision so that doctors and patients explore appeal the logic behind the prediction (Došilović and Hlupić,2018).

The increasing legal and ethical concerns have stressed the need for making AI explainable. The regulations like the General Data Protection Regulation (GDPR) in the European Union emphasize the right to explanation any AI model interpretable. It impacts human rights is involved in financial decisions (Samek et al., 2017).Transparency in AI is pivotal for bias mitigation and fairness, since opaque models unknowingly propagate discrimination if their decision-making process is not subjected to sufficient scrutiny (Ryo,2022). XAI include feature importance, local interpretable model-agnostic explanations Shapley additive explanations and visualizing the system through saliency maps. The complex ML models interpretable by providing insight into the way the predictions are generated, enabling better validation of the models and their ethical deployment towards AI (Ryo et al., 2021).

AI models like neural networks, ensemble learning tend to have better accuracy accompanied by a lack of interpretability. While training machine learning models, if you discover the best balance between accuracy and transparency. It would be the best approach that one could adopt for the usage of AI in the decision-making process. This study examines the contribution of XAI to ML interpretability, assesses XAI techniques, and discusses their practical approach in several sectors (Aslam et al., 2022).

1.1.1 Importance of Explainable AI

Artificial Intelligence refers to techniques and methodology that aim to make machine learning (ML) models more interpretable and transparent(Ali et al., 2023). Traditional black-box AIs offer you an accurate prediction, but you generally will have no idea how it reached this conclusion, and this has led XAI to be defined as a form of AI that extends to explain its model (Dwivedi et al.,

2023). XAI's main goals include increasing transparency, building up user trust, promoting model debugging, enabling ethical deployment and model validation and supporting regulatory compliance (Machlev et al., 2022). A primary motivation for XAI is to enhance trust in AI systems by offering intelligible rationale for model decisions.

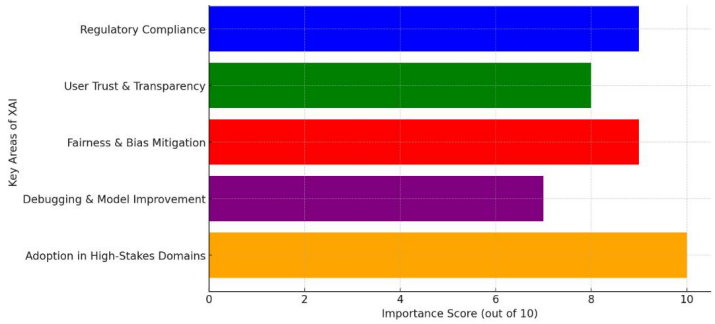
1.1.2 Interpretable Design: Utility in the Context of Regulatory Compliance, Trust and Fairness

Interpretability of ML models is important to ensuring compliance with regulations, trust, and fairness in AI applications. It is many international regulatory frameworks, like that of the General Data Protection Regulation in the European Union, mandate explaining AI-assisted decisions affecting individuals (Hong et al., 2020). The 'right to explanation' embedded in the GDPR requires organizations to provide humans comprehensible explanations of algorithmic outcomes, especially, but by no means only (Wachter and Floridi, 2017) in sensitive domains like financial credit scoring, hiring, or healthcare. The recent negotiation of the European Commission's proposed AI Act, which aims to establish centralized and transparent guidelines for high-risk AI systems, a legal push for explainable AI has been instigated (Lisboa et al., 2023).

XAI enables users to understand how and why models make certain decisions, building trust in AI systems. Trust is a cornerstone for AI systems to be adopted at scale, especially in areas where human life and livelihood are at stake. In the medical diagnostics setting (Nasarian et al. 2024) demonstrate that interpretable AI note that finders help physicians to investigate more in predictions made by AI systems, which, due to those helps to decrease uncertainty and make better evidence-based clinical decisions.

XAI plays a pivotal role in establishing fairness and reducing bias in the AI models. Since they are trained using biased training data or errors in the algorithm, some explicit models with black-box environments haphazardly amplify implicit societal biases. Explainability methods such as SHAP (LIME (Local Interpretable Model-Agnostic Explanations) assist organizations in detecting and addressing discriminatory tendencies in the outcomes of AI systems (Tursunalieva et al., 2024). Explainable AI will be the cornerstone to increase transparency, build trust among users, comply with particular regulations and guarantee the fairness of AI-powered solutions. XAI will play a significant role in helping to avoid this situation in the future because AI technologies are permeating innumerable areas of our lives.

Figure No.01: Key roles of Explainable AI in machine learning



1.2 Research Objectives

Investigate the role of XAI in enhancing ML interpretability.

Analyze various XAI techniques and their effectiveness.

Evaluate the trade-off between accuracy and explainability.

2 Literature Review

2.1 AI Interpretability and Explainability: Definitions

Explainability is one of the paradigms of AI that aims to back trace the reasoning process for the internally conducted processes that lead to decisions of the model into formats interpretable for humans (Linardatos et al., 2020). It is concerned with explaining why a model made the predictions it did. On the other hand, Interpretability is defined in terms of degree to which a human understands the cause of a given decision with respect to the input features of the model as it pertains to causality (Kamath and Liu, 2021). It is centered around explaining the reason of developed model and in most cases is evaluated against how clear and minimalistic the model is.

2.2 These definitions account for the various aspects of explainability and interpretability.

In provide definitions for these terms, explainability seeks to analyze complex models while operating at a higher level of abstraction. One the other hand, interpretability is concerned with a lower level, specific class of models that provide understandability from the outset (Kamath and Liu,2021).

2.3 The Difference Between Model-Specific and Model-Agnostic Explainability

Explainability approaches are further divided as either model-specific explainability, where more attention is paid to a single model, or model-agnostic explainability where specialized attention isn't required, for example, in range of machines. To begin with explainability, models try to explain themselves or their attempts to communicate using SHAP Additive Explanations tree boosted interpretations and even in certain cases the expansion of nurture trained Cams(Wang, 2024).

Techniques seek to scrutinize relations between inputs or outputs from the device or model. Partial Dependence Plots which depict the impact of parts or set features in relation to the predisposition of the designs are (Friedman, 200).These form part of the explanatory inclusions provided by Local Interpretable Model-agnostic Explanations that stand in as replacements or depend for the interpreted devices (Sandu and Trausan-Matu, 2022). Through model specific and multi sides explainability users are empowered and protected by certain tools ensuring precision AI applicability interpretations. This leads to the efficient application on places where gaps matter the most such as self-driving technology, finance, and healthcare systems (Knap, 2024).

2.4 The Evolution of Explainable AI

AI seeks to disclose the rationale behind every action an AI oracle takes. This branch of research was bred out of more advanced AI technologies that were utilized in sensitive domains such as healthcare or finance where possessing the ability to explain AI made a world of difference. XAI has progressed immensely, transforming from an area with limited research to a focus of the entire AI community (Goebel et al., 2018).

XAI research development began from the assumption that everything an AI does needs to be as transparent as possible. This assumption simplified interpreting results. AI

systems, including decision trees and rule-based systems were readily interpretable due to their simplistic nature. These deep learning systems were extremely performant but extremely opaque in terms of the reasoning behind their decisions, leading them to be dubbed "black-box" systems (Fernandez et al., 2019).

'Explainable AI' became popular in the late 2000s and early 2010s, especially when LIME (Ribeiro et al., 2016). These techniques made it possible to explain the predictions from black box models using approximations of the models with interpretable, simpler models. The more developed methods were created for improving model explainability, such as attention mechanisms in neural networks (Keneni et al., 2019). Defense Advanced Research Projects Agency launched the Explainable Artificial Intelligence program with the goal of advancing research in this field. (Wang et al., 2022).

2.5 Regulatory Framework and Ethical Issues Attention Areas

The growth of AI has brought regulatory frameworks and ethical considerations into play that may have hindered the implementation and growth of XAI technologies. It is a good thing that governments and regulatory bodies have begun to realize that AI models not only be useful and effective but fair, open, and accountable (Leenes et al., 2017).

An important step in this direction was made by the European Union with the arrival of the General Data Protection Regulation in 2018. The GDPR is one of several pieces of legislation of which the requirements for explainability in AI systems (most obviously given its focus on automated decision-making processes) form a part. GDPR, in Article 22, provides protection to individuals against being solely subject to decision-making based on automated profiling that affords them legal effects or similarly significantly affects them. This has created a greater demand for AI systems that account for decisions they produce (Abbas et al., 2014).

The law prescribes risk transparency and accountability in AI systems by mandating the justification of high-risk AI models to users. The AI Act has proposed that AI systems are constructed such that outputs are provided in an explainable way, in particular where the rights and freedoms of individuals are at stake (Iltis et al., 2023).XAI ethical concerns are not limited to adherence to legal norms. The AI explainability debate concerns bias, fairness, and discrimination. Discrimination tends to be further encoded in an unexplainable way in these models, making it practically infeasible to adjust the discrimination of the outputs they generate (O'Sullivan et al., 2019).

It is ethical considerations regarding data privacy and misuse of some approaches under the umbrella of explainable AI like over-guiding or dumbing down the justification for certain decisions, constitute a prickly concern in responsible development of AI. A balance talking about phenomena as well as policies, claiming to provide fairness with respect to such objectives while protecting privacy and accuracy, is a major challenge that researchers and policy enforcers face (Mbah, 2024).

2.6 Techniques for XAI in Machine Learning

Developments in the complexity of machine learning models have led to a demand for transparent and inspectable models. Several techniques have been developed to assist in explaining machine learning predictions, some of which are model type-specific and some are designed to be more global (Dwivedi et al., 2023).

2.7 Feature Importance Methods

Partial Dependence Plots illustrate the dependency of a predicted value in relation to a feature while keeping the rest of the features intact(König et al., 2021). PDPs enable analysis of the contributions of individual features to the predicted values, which is useful for capturing non-linearities (Altmann et al., 2010).

2.8 Model-Agnostic Approaches

These approaches enable the use of a framework irrespective of the type of model used for machine learning. The most popular methods that fall in this category include LIME, SHAP, and Anchors. LIME, a technique under local interpretable model agnostic explanations, uses an interpretable model to make predictions for complex "black-box" models (Ribeiro and Guestrin,2016). It is beneficial because it offers local and common interpretability and is very consistent in explaining predictions. "Anchors" is an approach that offers more comprehensible terms of explanation by spotting the so-called "anchors" that retain sufficient "rules" of a model decision and aid in its effortless interpretation (Viana et al., 2021).

The specific approaches deal with particular groups of models and therefore utilize their structure and mechanisms to provide explanations. With regard to decision trees, they are self-explanatory because their decision-making abilities represented as a diagram tree that branches out from basic values into more advanced features and surfaces. Each internal node represents a feature test, and each leaf node corresponds to a prediction (Molnar et al., 2020). Attention mechanisms in neural networks highlight which portion of the input data a neural network 'attends to' while predicting something. They place values on different sections of the input presented with the model, putting emphasis on the parts that have the most effect on the model's choice (Soleymani et al., 2022).

2.9 Visualization-Based Methods

Visualization-based methods offer non-technical approaches to interpreting a machine learning model by making visual representations of the model diagnostic. The two principal techniques in this respect are saliency maps and Grad-CAM. Saliency maps are useful to see which sections of a particular image influence the decision made by the model. Saliency maps weaken the image by modeling the gradients of the output with respect to the input and amplifying the regions of the image that are more definitive for prediction (Yang et al., 2019).

It generates Grad-CAM heatmaps that visualize the regions of the image that are primarily responsible for the model's decisions through the output's gradients in relation to the feature maps of the last convolutional layer. Grad-CAM allows for more refined visual interpretations of the

deep learning models which benefits applications like object detection (Li et al.,2023).

2.10 Applications of XAI in Various Sectors

The development of Explainable AI has transformed multiple domains by improving the trust, transparency, and accountability of AI-enabled systems. In some industries like healthcare, finance, law, or any field with high-stakes decisions, being able to explain how an AI model makes decisions is pivotal (Gadekallu et al., 2024).

2.11 Healthcare: Medical Diagnostics and Treatment Proposals

Deep learning and other AI models are very effective for tasks such as diagnosing diseases based on medical images, predicting outcomes for patients, and recommending treatment (Wennberg, 1984). These machines are often treated as black boxes by healthcare staff who do not necessarily comprehend the reasoning system used to arrive at particular diagnoses or treatment recommendations. Explainable AI techniques are used with medical AI systems to increase interpretability. XAI identify what features, such as patient history or specific biomarkers, were relevant in predicting a disease like cancer. It allows healthcare professionals to verify the justification of AI with medically relevant matters and through interpretation to determine the patient's treatment within medical correctness (Casal-Guisande et al., 2023).

2.12 Finance: Detection of Fraud and Scoring Credit

AI is widely used in fraud detection, credit scoring, and even algorithmic trading. But all these financial services need to comply with the compliance requirements of the GDPR of European Union and the Fair Credit Reporting Act (FCRA) of the United States. XAI helps in meeting these needs by enhancing the verifiability of information obtained by AI (Chen et al., 2020).This is important for fraud, for example, when you want to explain what factors were responsible for marking a transaction suspicious, such as the unusually high amount of transaction or additional outlier geographic locations. A great example would be SHAP values that help contribute towards breaking down what a model's decision looked like and the contribution that each feature has in the case of fraud detection. XAI techniques in credit scoring explain how different variables such as income levels, payment histories and debt-to-income ratios affect a person's creditworthiness (Al-Hashedi and Magalingam,2021).

2.13 AI and its Applications in Legal Decision Making

AI is progressively being applied more in the legal and criminal justice fields to enhance a variety of functions, including risk assessment, sentencing, and even legal research. Using an AI model, one estimates a convicted criminal's recidivism (reoffending) chances and help the judge to decide their optimum sentences (Scherer, 2019). AI systems seem advantageous in these instances; there is ambiguity in the equity, prejudice, and responsibility if the system makes such critical decisions on its own. XAI methods in the legal field seek to explain how AI models reach decisions relative to sentencing and risk assessment. AI models are tasked with predicting recidivism, LIME

help clarify which features, like prior convictions, employment status do the heaviest lifting. Models of this nature interpreted and scrutinized so that stakeholders monitor the application of ethical legal AI and minimize discrimination and bias in legal ramifications (Kabir and Alam, 2023).

It is the responsibility of AI to make decisions in real time within a dynamic, unpredictable, and incompletely observable environment which includes recognizing through sensors the presence of various objects and the existing traffic relations. Because of the safety-critical aspect of autonomous systems, it is equally important to know how AI models make these decisions, in what ways and for what reasons do they need to ensure the reliability and safety of their ethical behavior (Giuffrida, 2019).

Artificial Intelligence (XAI) are used increasingly in autonomous systems, with methods such as Saliency Maps and Grad-CAM being used to show the reasoning behind a machine's decision. For instance, in self-driving cars, XAI pinpoint the specific elements within the surrounding features, such as pedestrians, road signs, or other vehicles to pedestrians all of which the AI system used when executing operations. This aids engineers in understanding the specific aspects of the world the system's perception are based on and offers insight in the event of a crash or a system failure (Taruffo, 1998).

3 Methodology

3.1 Research Design

The writer adopts a mixed-methods approach through a literature review and empirical analysis to explore the role of XAI in increasing the interpretability of machine learning. The literature was reviewed of existing research, industry reports and regulatory frameworks to formulate a theoretical structure. In the empirical analysis, explainable AI techniques, including LIME, SHAP, and Grad-CAM, are applied to evaluate their effectiveness and the trade-offs between the two constructs, namely, accuracy and explainability. This help to improve the robustness of the inferences made, as it provides a multi-dimensional view of XAI impact by looking at qualitative information alongside quantitative measures.

3.2 Dataset and Model Selection

This research employs multiple machine learning approaches such as decision trees, random forests, neural networks, and deep learning to assess the efficacy of XAI methods. These models offer varying degrees of simplicity, enabling a trade-off between interpretability and predictive performance. These datasets useful to evaluate the performance of various XAI methods described at our method in a variety of domains and it would be a comprehensive measure for explainability in machine learning.

3.3 Explainability Techniques Evaluated

This paper is a preliminary study on the comparison of the most used XAI techniques, like SHAP, LIME, rule-based explanations, and saliency maps. They were selected for their ability to provide insight into a wide range of machine learning models, from decision trees to deep teaching networks. This approach enables the joint evaluation of XAI

techniques in the wider context of different examples of application.

3.4 Case Study Implementation

This study investigates the effect of XAI techniques on a real-life domain, such as health care, finance, or autonomous systems. When trained accordingly, a chosen model in machine learning will be analyzed using SHAP to foreshadow predictions using LIME and saliency maps in order to explain them more explicitly and with more transparency. Furthermore, a user perception study will be done to assess different health acknowledgment behaviors of stakeholders (e.g., doctors, financial analysts, or engineers) and examine how they perceive the usefulness and clarity of these explanations.

4 Results and Discussion

4.1 Evaluation of XAI Techniques

It compares the performance of SHAP, LIME, rule-based explanations and saliency maps for a diverse set of machine-learning models (decision trees, random forests, neural networks, and deep learning) and benchmark datasets (UCI datasets, MNIST, and COMPAS). It is important to note the results, as SHAP offers reliable and globally interpretable explanations that are practical for complex models like neural networks. LIME provides local interpretability but is stable only with less complex models deep learning makes LIME unstable. Rule-based explanations work better for structured data use cases like legal decision-making, where the need for transparency is paramount. Saliency maps are widely used in models relying on image information, as they validate the importance of features in the model but often sacrifice clarity for non-expert users. However, in general, the results highlight a trade-off of accuracy vs. interpretability, with simpler models (decision trees) explicable at the expense of accuracy and deep learning models needing advanced XAI procedures for information extraction.

XAI Technique	Interpretability Score (0-10)	Stability (0-10)	Computational Cost	Best Suited For
SHAP	9	8	High	Complex Models
LIME	7	5	Medium	Simple Models
Rule-based Explanations	8	9	Low	Structured Data
Saliency Maps	6	6	High	Image-based Models

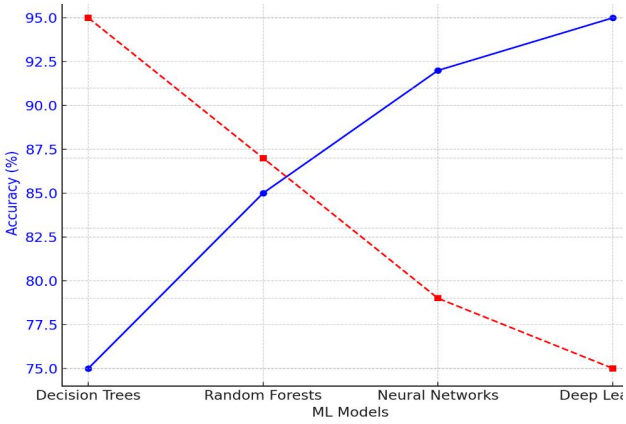
Table No.01: *statistical comparison table for different XAI techniques:*

4.2 Trade-offs Between Accuracy and Explainability

The paper explores the balance between performance and explainability on multiple machine learning models. Lite models such as decision trees and rule-based models give high interpretability but relatively lower accuracy; on the other hand, neural networks and deep learning are complex models. With such models, we achieve higher accuracy, but for explainability, we need to apply

advanced XAI techniques. For example, while SHAP and LIME make black-box models interpretable, they introduce computational complexity. On the other hand, decision trees and linear models are inherently explainable but may not perform best on complex datasets. Results indicate a trade-off between accuracy and discernability, considering that full explainability and high accuracy are challenging to achieve depending on the application and industry standards.

Figure No.02: Trade-off Between Accuracy and Explainability in ML Models



4.3 Impact on Stakeholders and Decision-Making

Explainability is important to build trust and encourage the adoption of AI-driven decision-making in multiple sectors. Healthcare workers, finance experts, lawyers, and engineers trust assertions from AI models when deciding matters of great importance. But without transparency, such users may be reluctant to trust or adopt AI recommendations completely. Interpretable AI models provide transparency that increases stakeholder trust, regulatory adherence and responsibility. In the health care vertical, an explainable model help doctors to understand the rationale behind a diagnosis which results in building trust with the patients. The explainable AI in credit scoring promotes fairness and mitigates bias concerns. AI interpretability means better stakeholder engagement and increased ethical AI deployment.

Stakeholder	AI Application	Impact of Explainability
Healthcare Professionals	Medical Diagnosis	Enhances trust in AI-assisted diagnoses and treatment recommendations.
Financial Analysts	Fraud Detection & Credit Scoring	Ensures transparency in credit decisions and fraud detection models.
Legal Experts	Legal Decision-Making	Improves fairness and accountability in AI-driven legal judgments.
Engineers	Autonomous Systems	Increases safety and reliability of autonomous systems and robotics.

Table No.03: the impact of explainability on stakeholders and decision-making:

4.4 Challenges in Implementing XAI

AI quite challenging in practice especially in terms of scalability, consistency, and real time application. Many XAI methods, like SHAP and LIME, demand high computational resources, and hence they hardly be applied to large models and big data settings, resulting in scalability issues. This is a crucial challenge to strike, especially as AI systems grow in size and complexity, maximizing explainability while minimizing potential

efficiency drops. Inconsistent explanations are another big one. Different XAI methods often provide different explanations for the same prediction, which result in contradictory explanations.

Certain explanations (like LIME ones) vary within multiple runs, which makes trusting AI decisions harder. Such variability means that it may not be possible to point to a singular definitive explanation. This is applicability in real-time is a major challenge, particularly in domains such as autonomous driving and fraud detection, where decisions made instantly. The explainability techniques add computational overhead, which slows down the AI model and is impractical for providing an explanation in real time. Researchers and developers pay attention to optimized algorithms, hybrid XAI approaches and efficient computational techniques that manage the trade-off between transparency and performance to surmount these challenges.

5. Conclusion

5.1 Summary of Findings

This study pays attention to the importance of Explainable AI, XAI, in addressing the challenge of interpreting Machine Learning, ML, models. The primary concern is that, on most circumstances, more advanced ML systems including, but not limited to neural and deep learning structures, are very accurate as a result which renders explanation techniques pivotal for trust, fairness, and accountability in AI-powered decisions. The examination of the other various techniques of XAI proved that SHAP and LIME are suitable for black-box models while rule-based and decision tree systems are much more accurate but have little to no productivity.

The study found the accuracy and explainability gap and points peculiarity which illustrates the need for an equilibrium between model efficiency and efficiency transparency. The research brought to light other challenges such as lack of modularity, conflicting explanations, and imposed hindrance to real-time use. In order to tackle these problems, more sophisticated methods of XAI and composite solutions need to be developed in order for AI to be usable and interpretable in different domains.

5.2 Practical Implications

Agencies and policymakers use Explainable AI to encourage ethical use of AI by ensuring fairness, transparency, and accountability in AI-based decision-making. With XAI technologies, companies build the confidence of stakeholders in high-risk industries such as healthcare, finance, and even criminal justice. The provision of explainability enables practitioners to check AI outcomes for biases, which ensures fairness. For policymakers, XAI offers an initial step towards meeting the internationally set standards of the EU AI Act and GDPR which underpin AI explainability. The use of explainable models assists organizations in complying with legislation for increased responsibility and improved public trust in the AI. An organization that goes along with the XAI approach will be able to create more user-friendly AI systems, thereby enhancing human-AI interaction. In the quest for ethical and high-performance AI,

organizations are better placed to design AI systems that are more understandable and interpretable.

5.3 Limitations of the Study

This research has multiple weaknesses connected to the choice of datasets, model applicability, and practical implementation. One primary limitation is the scarcity of appropriate datasets that used for the assessment of XAI evaluations. Most available datasets do not reflect real-world complexities which bias model explanations. The research considers several ML models; however, these outcomes may not be applicable to all AI-dependent sectors. .

XAI approaches may be efficient in environments with structured data such as finance, but they will have difficulties with unstructured data, including medical imaging or text documents. Lastly, the issues of practical implementation still persist. Most XAI techniques require significant computational resources, which compromises their real-time AI system integration. There is the issue of varying AI explanation interpretations by different stakeholders which compromises consistency in trust.

5.4 Future Research Directions

AI prioritize enhancing practical real-time XAI implementation, improving explanations at the user level and harmonizing XAI with governance structures for AI. An important opportunity is the creation of low-latency XAI methods that supply instant insight with minimal additional computation. This would permit more practical use of AI in autonomous systems and fraud detection. Another significant improvement stems from the ability to tailor user-centric explanations to different stakeholders.

This Research is needed to provide human-centered AI models that generate explanations that non-technical users, policymakers and lay industry professionals can easily understand. To ensure regulatory compliance and the ethical use of AI, integrating XAI into the governance structures of AI is critical. These explainability methods with existing global AI policies such as the EU AI Act and GDPR, would ensure that AI models are fair, transparent, and accountable. It is established in these systems and the development of responsible AI devoid of bias is facilitated.

References

s, R., Michael, K., & Michael, M. G. (2014). The regulatory considerations and ethical dilemmas of location-based services (LBS) A literature review. *Information Technology & People*, *27*(1), 2-20.

ıshedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, *40*, 100402.

., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion*, *99*, 101805.

ann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected

feature importance measure. *Bioinformatics*, *26*(10), 1340-1347.

Aslam, N., Khan, I. U., Mirza, S., AlOwayed, A., Anis, F. M., Aljuaid, R. M., & Baageel, R. (2022). Interpretable machine learning models for malicious domains detection using explainable artificial intelligence (XAI). *Sustainability*, *14*(12), 7375.

Casal-Guisande, M., Torres-Durán, M., Mosteiro-Añón, M., Cerqueiro-Pequeño, J., Bouza-Rodríguez, J. B., Fernández-Villar, A., & Comesaña-Campos, A. (2023). Design and conceptual proposal of an intelligent clinical decision support system for the diagnosis of suspicious obstructive sleep apnea patients from health profile. *International journal of environmental research and public health*, *20*(4), 3627.

Chen, K., Yadav, A., Khan, A., & Zhu, K. (2020). Credit fraud detection based on hybrid credit scoring model. *Procedia Computer Science*, *167*, 2-8.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018, May). Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210-0215). IEEE.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., ... & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), 1-33.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., ... & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), 1-33.

ández, A., Herrera, F., Cordon, O., del Jesus, M. J., & Marcelloni, F. (2019). Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational intelligence magazine*, *14*(1), 69-81.

kallu, T. R., Maddikunta, P. K. R., Boopathy, P., Deepa, N., Chengoden, R., Victor, N., ... & Dev, K. (2024). Xai for industry 5.0-concepts, opportunities, challenges and future directions. *IEEE Open Journal of the Communications Society*.

rida, I. (2019). Liability for AI decision-making: Some legal and ethical considerations. *Fordham L. Rev.*, *88*, 439.

el, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., ... & Holzinger, A. (2018). Explainable AI: the new 42? In *International cross-domain conference for machine learning and knowledge extraction* (pp. 295-303). Springer, Cham.

, S. R., Hullman, J., & Bertini, E. (2020). Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1-26.

A. S., Koster, G., Reeves, E., & Matthews, K. R. (2023). Ethical, legal, regulatory, and policy issues concerning embryoids: a systematic review of the literature. *Stem Cell Research & Therapy*, *14*(1), 209.

, M. S., & Alam, M. N. (2023). The role of AI technology for legal research and decision making. *Title of the Journal*.

ath, U., & Liu, J. (2021). Explainable artificial intelligence: an introduction to interpretable machine learning.

Kamath, U., & Liu, J. (2021). Explainable artificial intelligence: an introduction to interpretable machine learning.

Kamath, U., & Liu, J. (2021). Explainable artificial intelligence: an introduction to interpretable machine learning.

Keneni, B. M., Kaur, D., Al Bataineh, A., Devabhaktuni, V. K., Javaid, A. Y., Zaientz, J. D., & Marinier, R. P. (2019). Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. *IEEE Access*, *7*, 17001-17016.

Knap, M. (2024). Model-agnostic XAI models: benefits, limitations and research directions.

König, G., Molnar, C., Bischl, B., & Grosse-Wentrup, M. (2021, January). Relative feature importance. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 9318-9325). IEEE.

Leenes, R., Palmerini, E., Koops, B. J., Bertolini, A., Salvini, P., & Lucivero, F. (2017). Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues. *Law, Innovation and Technology*, *9*(1), 1-44.

Li, X., Wu, K., & Liang, Y. (2023). A Review of Agricultural Land Functions: Analysis and Visualization Based on Bibliometrics. *Land*, *12*(3), 561.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Lisboa, P. J., Saralajew, S., Vellido, A., Fernández-Domenech, R., & Villmann, T. (2023). The coming of age of interpretable and explainable machine learning models. *Neurocomputing*, *535*, 25-39.

ılev, R., Heistrene, L., Perl, M., Levy, K. Y., Belikov, J., Mannor, S., & Levron, Y. (2022). Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, *9*, 100169.

ı, G. O. (2024). The Role of Artificial Intelligence in Shaping Future Intellectual Property Law and Policy: Regulatory Challenges and Ethical Considerations. *Journal homepage: www. ijrpr. com ISSN*, *2582*, 7421.

ar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., ... & Bischl, B. (2020, July). General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers* (pp. 39-68). Cham: Springer International Publishing.

rian, E., Alizadehsani, R., Acharya, U. R., & Tsui, K. L. (2024). Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework. *Information Fusion*, 102412.

livan, S., Nevejans, N., Allen, C., Blyth, A., Leonard, S., Pagallo, U., ... & Ashrafian, H. (2019). Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *The international journal of medical robotics and computer assisted surgery*, *15*(1), e1968.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Ryo, M. (2022). Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. *Artificial Intelligence in Agriculture*, *6*, 257-265.

Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2021). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, *44*(2), 199-205.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Sandu, M. G., & Trausan-Matu, S. (2022). Comparing model-agnostic and model-specific XAI methods in Natural Language Processing. In *RoCHI* (pp. 115-118).

Scherer, M. (2019). Artificial Intelligence and Legal Decision-Making: The Wide Open? *Journal of international arbitration*, *36*(5).

Soleymani, M., Ali, R. E., Mahdavifar, H., & Avestimehr, A. S. (2022, June). ApproxIFER: A model-agnostic approach to resilient and robust prediction serving systems. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 8, pp. 8342-8350).

Taruffo, M. (1998). Judicial decisions and artificial intelligence. In *Judicial Applications of Artificial Intelligence* (pp. 207-220). Dordrecht: Springer Netherlands.

Tursunalieva, A., Alexander, D. L., Dunne, R., Li, J., Riera, L., & Zhao, Y. (2024). Making sense of machine learning: a review of interpretation techniques and their applications. *Applied Sciences*, *14*(2), 496.

a, C. M., Santos, M., Freire, D., Abrantes, P., & Rocha, J. (2021). Evaluation of the factors explaining the use of agricultural land: A machine learning and model-agnostic approach. *Ecological Indicators*, *131*, 108200..

g, Y. (2024). A comparative analysis of model agnostic techniques for explainable artificial intelligence. *Research Reports on Computer Science*, 25-33.

g, Y., Liu, W., & Liu, X. (2022). Explainable AI techniques with application to NBA gameplay prediction. *Neurocomputing*, *483*, 59-71.

nberg, J. E. (1984). Dealing with medical practice variations: a proposal for action. *Health affairs*, *3*(2), 6-33.

, H., Shao, X., & Wu, M. (2019). A review on ecosystem health research: A visualization based on Cite Space. *Sustainability*, *11*(18), 4908.