ISSN : 2350-0743

*International Journal of Recent Advances in Multidisciplinary Research*

*Vol. 11, Issue 12, pp. 10531-10536, December, 2024*

# RESEARCH ARTICLE

## EXPLAINABLE AI IN ML: THE PATH TO TRANSPARENCY AND ACCOUNTABILITY

***Govindaiah Simuni**

Vice President, Technology Manager, Bank of America, Charlotte, NC, USA

**ARTICLE INFO**

***Article History***

Received 20th September, 2024
Received in revised form
16th October, 2024
Accepted 27th November, 2024
Published online 29th December, 2024

***Keywords:***

Explainable Artificial Intelligence (XAI), Machine Learning Interpretability, Predictive Maintenance, Healthcare AI Applications, Transparency and Accountability in AI.

****Corresponding author:***
***Govindaiah Simuni***

**ABSTRACT**

This paper introduces Explainable Artificial Intelligence (XAI) as a requisite solution for increasing the interpretability and trustworthiness of ML systems. This paper explores the importance of XAI through case studies in two key sectors: manufacturing and healthcare. The first case relates to a predictive maintenance application that uses XAI to anticipate the likelihood of machinery failure using a gradient-boosting decision tree, which provides detailed recommendations for optimizing productivity. The second case study is based on the healthcare sector. LIME and SHAP tools for AI model explanation enhance trust in the effectiveness of diabetes predictions by AI-aided medical diagnosis. Both cases prove that XAI helps better understand the workings of exhibited machine learning models and aligns with ethical decision-making and regulatory requirements. Therefore, in the last part of the paper, the authors discuss some of the drawbacks of current forms of XAI and potential advancements in this field, including horizontality and better incorporation into the systems that use AI. This discovery supports using XAI to foster more transparent, responsible, and trustful AI applications in numerous industries.



***Citation: Govindaiah Simuni. 2024.*** *"Explainable AI in Ml: The path to Transparency and Accountability", International Journal of Recent Advances in Multidisciplinary Research,* 11, (12), 10531-10536.

# INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) are progressing to be the main drivers of innovation and changes across many industries, including healthcare and finance. AI Technologies facilitate the use of large data sets and processes algorithms to identify trends, forecasts or aid in decision support systems [1]. With the advances in AI and ML the usage has gone beyond automation to intricate problem-solving situations bringing changes and improvement to the various sectors and business. However, as the models become complex, the decision making process that the model possesses does not easily lend itself to understanding rendering them what is known as black box models. This is quite a problem as those who depend on these technologies will struggle to comprehend how decisions are made. Fig. 1 shows that the explainability can be applied in two stages: The first two time references prior to the Modelling and after the Modelling.

**Problem Statement:** The opaqueness associated with black-box models presents substantial problems of trust, responsibility, and ethics with users such as investors and consumers.

Whenever users cannot understand how certain predictions or decisions are made by the system, they start to question the credibility of the system. This is evident in critical usage such as patient treatment or financial investment where AI NS can significantly disrupt economy and health. Failure to grasp these two processes can result in adverse effects that include decision-making prejudice and risk from wrong decisions. Additionally, the legal requirements and the public consciousness have been growing more alert to the need for better responsibility and accountability within the algorithms and hence, increased transparency.

**Research Purpose:** To these challenges, Explainable AI (XAI) has been presented as an essential solution. XAI aims at increasing explainability of AI systems so it becomes easier for developers as well as the users to understand why AI has made certain decision. This approach is useful for a two-fold purpose: it indicates how the algorithms work and also explains why the particular prediction was made. The primary goals of XAI are to enhance interpretability, trust, and compliance with legal and ethical expectations from AI for users and to solve the user's problems. The key objectives of XAI include improving transparency, fostering user trust, and

ensuring that AI operates within ethical and legal frameworks while addressing users' needs effectively. By demystifying the decision-making processes, XAI aims to create a more responsible AI landscape where users can engage with AI systems confidently.

**Importance of the Study:** This paper aims at discussing the possibility of XAI as an able tool for bringing increased transparency to the decision-making process, which is critical for high-risk application areas like healthcare and finance. From another perspective, the value of this work is in actualizing the creation of such universal guidelines and reference models that will emphasize the explainability of artificial intelligence tools. Stressing the importance of establishing trust and reliability of the AI systems, this research focuses on the value that the XAI brings to the process of understanding and interpreting decisions made by AI-based systems. Lastly, it is intended to demonstrate that incorporating XAI principles is a way to promote the deployment of more ethical AI in supporting society and minimizing certain concerns related to decision-making within black box models.Fig. 2 shows that the Explainability can begin with the categorisation by which there is the Model-Specific explainability, Model-Agnostic explainability, Model-Centric explainability Data-Centric explainability.

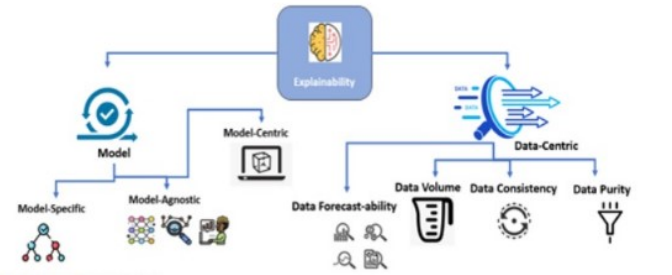


**Fig. 1. Explainability Techniques [12]**

**Research Aims and Objectives:** To explore how Explainable Artificial Intelligence (XAI) enhances transparency and accountability in machine learning for healthcare and finance.

Evaluate effective XAI techniques. Analyze case studies in healthcare and finance. Assess XAI's impact on stakeholder trust. Recommend best practices for XAI integration.

**Research Questions**

- How does Explainable AI (XAI) improve transparency and accountability in machine learning models?
- What are the benefits of XAI tools in predictive maintenance and healthcare decision-making?

## LITERATURE REVIEW

**Explainable AI: Definitions and Concepts:** XAI deals with techniques and methods that help make or explain the decisions of AI models understandably [2]. While black-box models are useful in making precise forecasts and converting them into solutions that people can understand, there is no explanation as to why these solutions were arrived at in the first place – this is where XAI looks to make a difference in an attempt to develop models that people can comprehend and go through the processes that lead to the suggested outcome. XAI techniques are generally divided into two categories: global explanations, which are more relevant to give the big picture of the model, and local explanations, which are based on the specific prediction or decision made by the model [3].

**Challenges of Transparency in Machine Learning:** In machine learning, the models are mostly either very large and very complex, especially the deep learning algorithms that are commonly used and that give rise to the black box, or mostly uninterpretable models where the primary challenge with such models is mostly unexplained decisions; thus, the need to give an understandable and justifiable decision, especially for high-risk industries such as health and finance [4]. If not for transparency, these models are more prone to these biases, meaning that the result of an automated process can be unfairly or unethically skewed. Moreover, legal requirements such as the GDPR already require decision-making AI systems to be explainable to increase accountability [5]. The Fig. 3Create the possibility to use some data for the training of an model in a particular learning procedure. What is obtained after this learning process is a learned function where inputs can be fed to it and it will give a prediction in an specific interface which is what the final user interacts with.

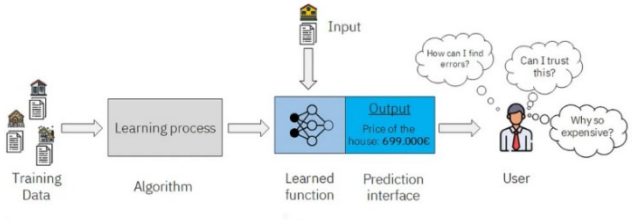


**Fig. 2. ML Workflow [11]**

**Existing XAI Techniques and Frameworks**

**Local Interpretable Model-agnostic Explanations (LIME):** LIME makes individual predictions understandable by creating locally interpretable models in the neighborhood of the specific prediction [6].

**Shapley Additive Explanations (SHAP):** SHAP associates each feature employed in a model's prediction with a contribution value, thus letting us understand which features were most critical in a prediction [7].

**Counterfactual Explanations:** These explanations concentrate on the description of how the prediction would be affected if some features are changed, and this assists users in identifying what may affect the result [8]. Altogether, these techniques help to make the process more transparent. However, there are difficulties in taking them further, in questions of their efficiency as a computational model and their capacity to represent real-world situations. For instance, semantics like LIME and SHAP may be time-consuming to compute and may not be efficient when applied to large, complex datasets. In fig. 4, shows that in the new and different learning process data is used to learn a function associated to an explainable model. What this means is that this function is capable of not only delivering the prediction but is also capable of explaining that prediction.

A new interface for explanation enlarges on some information that might allow for some understanding of why such a prediction was made.

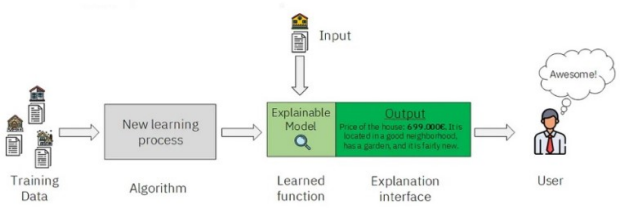


**Fig. 3. Explainable AI incorporated to the workflow [11]**

**The Significance of XAI in Little Applications:** Thus, the importance of openness can hardly be overestimated within rather serious and consequential spheres. In healthcare, AI models play the role of diagnosing diseases, treating them, prescribing them, and discovering new drugs. However, this comes with the drawback that anchoring could cause medical professionals not to trust the model's output if there is no explanation of how the diagnosis or recommendation is arrived at. Likewise, in finance, credit scoring models establish that loan approvals are possible, and if biases are left unaddressed, they nurture unfair credits [9]. In both fields, XAI serves the important purpose of removing AI from questionable and unlawful territory while ensuring users' trust [10]. In Fig. 5 XAI has been introduced in the ML life cycle, and it is taking the responsibility and to seal the expectation to explain and translate black-box algorithms, which are used for stakeholder critical business decision-making process; it becomes to increase their adoption and alignments.

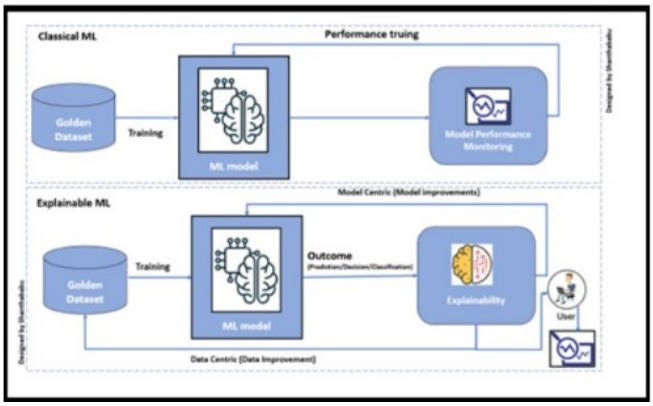


**Fig. 4. Block diagram of the Classical and XAI approaches [12]**

# METHODOLOGY

**Case Study Analysis:** This research employs a case study analysis methodology to explore the role of Explainable Artificial Intelligence (XAI) in enhancing transparency and accountability in machine learning applications. The study involves selecting relevant case studies from healthcare and finance sectors, where XAI techniques have been implemented to address the challenges posed by black-box models. Data will be collected through a comprehensive review of existing literature, including academic papers, industry reports, and real-world applications of XAI. Each case study will be analyzed to evaluate the effectiveness of different XAI methods, such as LIME and SHAP, in improving interpretability and fostering trust among stakeholders. This approach will facilitate a deeper understanding of the practical implications of XAI, providing insights into best practices for its integration in critical decision-making processes.

**Case Study Analysis**

**Case study [13]:** This paper introduced the concept of XAI in a special industrial PdM use case for manufacturing. With predictive maintenance, AI helps identify machines or tools that are about to fail so that they can be rectified, preventing them from contributing to operational downtime costs. In the present study, a machine learning model was designed with a gradient-boosting decision tree to fully take advantage of the efficiency of structured data. In order to predict when a machine might go wrong, the model was built from generalized production parameters. As the data set was unrealistic, the experimental approach allowed an understanding of how PM could be implemented in realistic manufacturing contexts. By incorporating XAI into the model, the engineers were able to determine the causes of the failure of the machines, consequently enabling them to make the right decisions concerning the early schedule for maintenance activities. XAI revealed broken machines, their age, usage patterns, and environmental conditions through features important for breakdown recognition. Such a focused method of analysis allowed maintenance teams to better identify priority areas, making maintenance more effective and reliable. From this particular case study based on the exhibition, it is clear that with explainable AI, there is a way in which such a system can boost the reliability of the predictions of the maintenance tasks, hence improving the overall acceptance of automated decision-making systems due to high transparency exhibited by explainable AI.

**Case study [14]:** The second case study aims to investigate how XAI can be used in healthcare, with more emphasis on medical diagnosis and the view of predicting if a given patient is diabetic. The models themselves are often opaque by design, a core tenet of the ML process, especially in critical areas of human activity such as healthcare, where misclassification can be costly. This work employed the ELI5 toolkit and LIME and SHAP frameworks to make the diabetes prediction model developed from clinical trial data interpretable. The ML model, written in Python, used patient health data to classify diabetes, and the XAI techniques then explained why certain classifications occurred.LIME and SHAP helped show all the variables in the predictive model, such as blood glucose level, age, and BMI. The interpretability enabled clinicians and researchers to clearly understand the decisions made by the model, and hence, they gave confidence to the AI outputs in the clinical context. The capacity to justify such decisions also provides for governance in the application of AI for medical purposes. Furthermore, XAI was also useful for drawing out patterns and key parameters that can be seen in the data and provided suggestions for clinicians and healthcare professionals. Thanks to XAI, which helped reveal the mechanisms of functioning of the ML model, the work of medical personnel became more informed as machine learning algorithms were clarified and patient outcomes became more positive. As evidenced in this paper, bringing the notion of XAI into use in the healthcare sector can substantially improve

the interpretability of the ML models and increase users' confidence in the models.

**Data Analysis**

**How the Data Will Be Analyzed Within the PESTEL Framework: The data collected from the articles will undergo rigorous analysis within the PESTEL framework to identify and evaluate the key factors influencing the adoption and effectiveness of AI and ML in cybersecurity:**
**Political Analysis:** A study on how government policies, regulations, and international organizations affect AI and ML in cyber defense measures. This involves the analysis of cybersecurity legislation, government funding programs, and geopolitical aspects that influence global cybersecurity plans.

**Economic Analysis:** Budget allocation for acquiring funds needed for AI and ML to be adopted in cybersecurity, economic comparative analyses between AI solutions and conventional methods, and policies that encourage organizations to embrace AI and ML in cybersecurity.

**Social Analysis:** The perception of society on AI and ML in relation to cybersecurity. This area would encompass public perception of AI and ML towards cybersecurity, ethical issues on AI and cybersecurity, impacts of artificial intelligence on the workforce regarding job loss, and requirements of talent for cybersecurity jobs, among other aspects of AI-influenced cybersecurity solutions**.**

**Technological Analysis:** Assessment of emerging technologies in AI, ML, and cybersecurity, such as the elucidation of new enhancements in AI algorithms, increased computation power, and the use of Artificial intelligence-based automation in numerous threat identification and response systems.

**Environmental Analysis:** There is general concern about environmental sustainability practices in the AI and ML technologies currently employed in cybersecurity, such as the power draw of AI boards/ chips and the footprint of data centers that support AI-intense cybersecurity solutions.

**Legal Analysis:** Explaining the existing legal standards in the field of AI/ML concerning cybersecurity, data protection relevant to the present legal frameworks, legal responsibility in conjunction with AI-enabled decision-making in the context of cyberattacks, and finally, the principles of intellectual property regarding technological developments including AI in the area of cybersecurity. Therefore, adopting the PESTEL analysis for these dimensions, this study seeks to give a comprehensive perspective of the external forces that impact the application, integration, and performance of AI and ML technologies in cyberspace. The research will provide rich information for researchers, policymakers, and ICT and security professionals who wish to implement better risk analysis and management techniques using AI and ML for advanced cybersecurity.

# DISCUSSION

**Comparison Across the Cases:** The two cases highlight how explainable AI is applicable across industries and why it is essential. Their application in production: XAI was used in the manufacturing field where predictive maintenance was the main component, and the model had to be transparent to perfectly forecast the failures in the machines. XAI was crucial in identifying critical maintenance tasks that would have consumed much time and resources in the event, therefore saving. On the other hand, in the healthcare case study, XAI aimed at predicting the likelihood of diabetes, and here, the capability of explaining the decisions made by the model had repercussions on the immediacy of trust with the clinicians.

In both cases, XAI was found to be central to providing interpretations and guaranteeing responsibility, although the objectives may have been different. For manufacturing, the main themes were related to achieving more productivity or optimally utilizing production time. In contrast, for the healthcare sector, the main themes were related to utilizing clinical decision support and consequently improving the patient's condition. However, common to both was the level of confidence individuals had in AI-driven predictions, where possible through the XAI approach.

**Significance to the Development of AI:** From these successful case studies, it is clear that the incorporation of the XAI into AI systems with large applications is necessary where more accountability is required in substantive industries. Practical applications of predictive models in different fields such as manufacturing and healthcare are now possible through data, and with XAI, models cannot work obscure and mysterious. As a result, the end-users of the AI models – the engineers or the medical personnel – can easily understand what the models are doing. Thus, it is easy for them to make proper decisions based on the output generated by the model.Furthermore, as legal frameworks pertaining to AI advance, XAI provides a route to meeting the lawful demands of different governing legislations, such as the GDPR, which requires AI to be opaque. The offering of clarity has the additional benefit of not only increasing confidence in AI systems but also increasing their use for proper purposes only.

**Challenges to present approaches of XAI:** However, issues are still evident even after presenting the successes of both the case studies presented here about the implementation of XAI. Most of the current XAI techniques include LIME, SHAP, and others, for example, and could be computationally expensive, especially when dealing with large data sets or complex models. Besides, these tools help to provide some insights, but they are not perfect, and the interpretations of such tools can be seriously misleading in highly complex circumstances. There is still much potential to improve their stability and apply them across various industries and model types, and future studies must consider honing these methods.Simply displaying sample products does not ensure that customers can identify each one or that they will complete a purchase when desired.

# CONCLUSION

**Summary of the key findings:** This paper has explored the role of Explainable AI (XAI) in enhancing transparency and accountability in machine learning systems, focusing on two

distinct case studies: predictive maintenance in manufacturing industries and diagnosis in the healthcare Industry. The first strategic application example illustrated how XAI methods can identify breakdown precursors to give maintenance staff visibility on machine risk of failure and enhance operational performance. The second case analysis also emphasized explainability in healthcare – when LIME and SHAP were used to elucidate diabetes predictions so clinicians could trust AI-driven decisions. It should be noted that both cases underline the necessity of transparency as a key achievement of effectiveness in utilizing AI technology and correcting it in conditions that can present a certain risk. When using XAI to explain the machine learning models, the world between autonomous AI decision-making and human comprehension is closing, encouraging correct and accountable actions.

**Major Findings Emerging from Case Studies and Literature on XAI**

**Transparency and Trust:** XAI enhances confidence in AI models by improving model interpretability, especially in fields such as healthcare and manufacturing.

**Operational Efficiency:** Applying XAI in the manufacturing industry helps explain how to perform maintenance, decrease machine downtime, and plan resources.

**Clinical Decision Support:** In healthcare, explainability means that doctors understand the basis for Artificial Intelligence's medical predictions well enough to trust them and provide the best outcome to their patients.

**Accountability and Ethics:** Through XAI, the ethical issues with adopting AI can be addressed by providing reasons for certain decisions made to enable organizations to meet regulatory requirements such as the GDPR.

**Future Research Directions:** As XAI continues to evolve, several opportunities exist for advancing research in this field:

**Scalability of XAI Techniques:** Future development should continue with the present methods of explanation, like LIME or SHAP, and adapt the present techniques, which are good but valid for rather small datasets.

**Integration with More Complex Models:** There is relatively little work on extending XAI techniques so that they apply to deep learning models and other highly complex models that are still largely opaque.

**Industry-Specific Applications:** Further studies should also investigate the possible extensions of XAI to other strategic domains, such as finance, autonomous systems, and cybersecurity, among others, in which explainability and, therefore, trustworthy AI systems are essential.

**Ethical AI Frameworks:** Consequently, as AI solutions are integrated and applied increasingly widely, XAI is capable of contributing to the formation of ethical AI guidelines and standards that would promote the use of optimally accurate but also fare, objective, and comprehensible AI systems. Addressing the future challenges and opportunities highlighted in this study will allow XAI to continue being the leading force in enhancing accountable and responsible AI development across the vast spectrum of its applications.

# REFERENCES


Khan M. M. and Vice, J. "Toward Accountable and Explainable Artificial Intelligence Part One: Theory and Examples," *IEEE Access*, vol. 10, pp. 99686–99701, 2022, doi: https://doi.org/10.1109/access.2022.3207812.

Islam, S. R. W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable Artificial Intelligence Approaches: A Survey," *arXiv.org*, 2021. https://arxiv.org/abs/2101.09429 (accessed Sep. 27, 2024).

Barredo Arrieta A. *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion*, vol. 58, no. 1, pp. 82–115, Jun. 2020, doi: https://doi.org/10.1016/j.inffus.2019.12.012.

Bunn, J. "Working in contexts for which transparency is important," *Records Management Journal*, vol. ahead-of-print, no. ahead-of-print, Apr. 2020, doi: https://doi.org/10.1108/rmj-08-2019-0038.

Gohel, P. P. Singh, and M. Mohanty, "Explainable AI: current status and future directions," *arXiv:2107.07045 [cs]*, Jul. 2021, Available: https://arxiv.org/abs/2107.07045

Ahmed, I. G. Jeon, and F. Piccialli, "From Artificial Intelligence to eXplainable Artificial Intelligence in Industry 4.0: A survey on What, How, and Where," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 1–1, 2022, doi: https://doi.org/10.1109/tii.2022.3146552.

Rawal, A. J. Mccoy, D. B. Rawat, B. Sadler, and R. Amant, "Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges and Perspectives," *IEEE Transactions on Artificial Intelligence*, pp. 1–1, 2021, doi: https://doi.org/10.1109/tai.2021.3133846.

Linardatos, P. V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: https://doi.org/10.3390/e23010018.

Md. Hasib, K. F. Rahman, R. Hasnat, and Md. G. R. Alam, "A Machine Learning and Explainable AI Approach for Predicting Secondary School Student Performance," *IEEE Xplore*, Jan. 01, 2022. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9720806

Ryo, M. "Explainable artificial intelligence and interpretable machine learning for agricultural data analysis," *Artificial Intelligence in Agriculture*, Nov. 2022, doi: https://doi.org/10.1016/j.aiia.2022.11.003.

"Explainable Artificial Intelligence," https://towardsdatascience.com, 2020. https://towardsdatascience.com/explainable-artificial-intelligence-14944563cc79

Pandian, S. "Explainable Artificial Intelligence (XAI) for AI & ML Engineers - DataScienceCentral.com," Data Science Central, Oct. 16, 2023. https://www.datasciencecentral.com/explainable-artificial-intelligence-xai-for-ai-ml-engineers/

Hrnjica B. and S. Softic, "Explainable AI in Manufacturing: A Predictive Maintenance Case Study," IFIP Advances in Information and Communication Technology, pp. 66–73, 2020, doi: https://doi.org/10.1007/978-3-030-57997-5_8.

Vishwarupe, V. P. M. Joshi, N. Mathias, S. Maheshwari, S. Mhaisalkar, and V. Pawar, “Explainable AI and Interpretable Machine Learning: A Case Study in Perspective,” Procedia Computer Science, vol. 204, pp. 869–876, 2022, doi: https://doi.org/ 10.1016/j. procs. 2022.08.105.

**Acronyms**

1. **XAI** - Explainable Artificial Intelligence
2. **ML** - Machine Learning
3. **AI** - Artificial Intelligence
4. **LIME** - Local Interpretable Model-agnostic Explanations
5. **SHAP** - Shapley Additive Explanations
6. **GDPR** - General Data Protection Regulation
7. **PdM** - Predictive Maintenance

*******