World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

Enhancing the transparency of data and ml models using explainable AI (XAI)

Vinayak Pillai *

The department of Information systems and operations management, The University of Texas at Arlington, Arlington, Texas, United States of America.

World Journal of Advanced Engineering Technology and Sciences, 2024, 13(01), 397–406

Publication history: Received on 07 August 2024; revised on 18 September 2024; accepted on 20 September 2024

Article DOI: https://doi.org/10.30574/wjaets.2024.13.1.0428

Abstract

To this end, this paper focuses on the increasing demand for the explainability of Machine Learning (ML) models especially in environments where these models are employed to make critical decisions such as in healthcare, finance, and law. Although the typical ML models are considered opaque, XAI provides a set of ways and means to propose making these models more transparent and, thus, easier to explain. This paper describes and analyzes the model-agnostic approach, method of intrinsic explanation, post-hoc explanation, and visualization instruments and demonstrates the use of XAI in various fields. The paper also speaks about the requirement of capturing the accuracy and interpretability for creating responsible and ethical AI.

Keywords: Explainable AI (XAI); Model Transparency; Machine Learning Interpretability; Data-driven Decision-Making; AI Ethics; Model-Agnostic Techniques

1. Introduction

1.1. Background and Motivation

Layman Explanation: Suppose you have a box that is able to foretell for instance, whether a person will get a loan from a certain bank or that a certain patient is at risk of a certain disease. It also means that you don't have any understanding of what information is fed to the box, how it reaches conclusions, and why it is wrong occasionally. This is somewhat of how some of the current AI systems function. They are very strong and effective tools that can produce precise forecasts, yet, in the majority of cases, we do not know what the principal mobilizing factors are. This lack of understanding can be a big problem particularly when such predictions impact peoples' lives.

Professional Insight: The advancement of artificial intelligence and machine learning techniques has made them popular across all industries including healthcare and finance. These technologies if deployed effectively have the capability of transforming the decision-making paradigms by offering insights from big data. However, with their increasing complexity, the models' decision-making processes are not easily traceable, hence the black box problem. The drive for this research comes from the increasing necessity to explain such models further and make their decisions more comprehensible to end users, regulators, and other parties.

1.2. The Black Box Problem in Machine Learning

Layman Explanation: A "black box" in AI is like a mysterious machine—you put information in, and it gives you an answer, but you have no idea how it got there. For example, if a computer says you don't qualify for a loan, it's important to know why. Did it look at your income? Your credit history? If we don't understand the reasons, it's hard to trust the decision or fix any mistakes.

* Corresponding author: Vinayak Pillai

Professional Insight: The term "black box" in the context of ML refers to models whose internal workings are not easily interpretable. Deep learning models, for instance, consist of numerous layers of interconnected nodes, making it challenging to trace how specific input data leads to a particular output. This lack of transparency poses significant risks, especially in critical applications where understanding the reasoning behind a decision is as crucial as the decision itself. The black-box nature of these models can lead to issues such as bias, unfair treatment, and difficulty in troubleshooting errors.

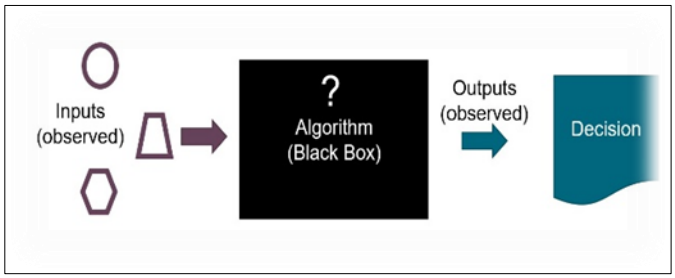

Figure 1 TYPE BIAS and ALGORITHM

1.3. The Importance of Transparency

Layman Explanation: In the context of AI, transparency refers to the ability to break the veil and look at what is happening behind the ‘‘black box.’’ It is rather like looking at a recipe in order to figure out which ingredients were used in baking a cake and all those other delightful things that we wouldn’t want to know about. If we could understand how an AI system arrives at its decisions then we can trust it more, we are able to debug it and ensure that it is not being prejudiced against anybody.

1.4. Explainable AI also known as XAI

Layman Explanation: XAI in short can be defined as an AI that has a guide to explain how it is getting to its results or answers. It shows you why the AI decided to make that decision and what it took into account in the course of reaching the conclusion. For instance, if an AI assigns a high probability of the person developing diabetes, XAI can explain that this decision was made based on the person’s diet, level of physical activity, and history of diabetes in first-degree relatives.

1.5. Challenges in Achieving Transparency

Layman Explanation: Making AI transparent isn’t always easy. Sometimes, the more accurate a model is, the harder it is to understand. It’s like trying to explain how a very complicated machine works—sometimes, the details are so complex that they’re hard to put into simple words. Plus, different people might need different kinds of explanations. A doctor and a patient might want to know different things about how an AI system makes decisions.

Professional Insight: Achieving transparency in ML models presents several challenges. One major challenge is the trade-off between accuracy and interpretability; simpler models are often more interpretable but may not perform as well as more complex models. Another challenge is the diversity of users and stakeholders, each of whom may require different types and levels of explanation. Furthermore, there is the issue of scalability, as some XAI techniques can be computationally intensive and may not be feasible for large-scale models. Finally, there is the risk of oversimplification, where explanations may not capture the full complexity of the model, potentially leading to misunderstandings.

1.6. Objectives of the Research

Layman Explanation: The goal of this research is to find better ways to make AI systems more understandable. We want to explore different tools and methods that can open up the "black box" of AI, making it easier for people to see how decisions are made and ensuring those decisions are fair and reliable.

2. Literature review

The literature review section is essential for grounding your research in the existing body of knowledge. It synthesizes the work that has been done on Explainable AI (XAI) and its application in enhancing the transparency of machine learning (ML) models. Below, we break down the literature review into subtopics, providing in-depth explanations that are both professional and accessible to a lay audience.

2.1. Historical Overview of Transparency in AI

Layman Explanation: The concept of transparency in AI isn't new. Early AI systems were simpler and easier to understand. Imagine a decision tree—it's like a flowchart that shows you every step leading to a decision. But as AI grew more complex, with systems like deep learning, it became harder to see inside and understand what was happening. This shift led to a focus on making these complex models more transparent.

Professional Insight: Early AI models, such as decision trees and rule-based systems, were inherently interpretable, allowing users to trace the decision-making process. As the field evolved, more complex models, such as deep neural networks and ensemble methods, emerged, offering superior performance but at the cost of interpretability. This historical shift from transparent to opaque models has spurred interest in developing methods to regain transparency without compromising model performance.

2.2. Model-Agnostic vs. Model-Specific Techniques

Layman Explanation: There are two main ways to make AI models more understandable. One way works for any type of model—these are called "model-agnostic" techniques. Think of them as a universal tool that can explain different kinds of AI systems. The other way is specific to certain types of models—these are "model-specific" techniques, which are like custom tools designed for a particular kind of AI.

Professional Insight: Model-agnostic techniques, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive exPlanations), offer flexibility as they can be applied to any ML model. These methods generate explanations that help users understand individual predictions or the overall model behavior. In contrast, model-specific techniques are tailored to particular types of models. For instance, decision trees and linear models have built-in interpretability, while neural networks may require specialized techniques like saliency maps or feature visualization. The literature suggests that both approaches have their merits, depending on the context and the specific needs of the application.

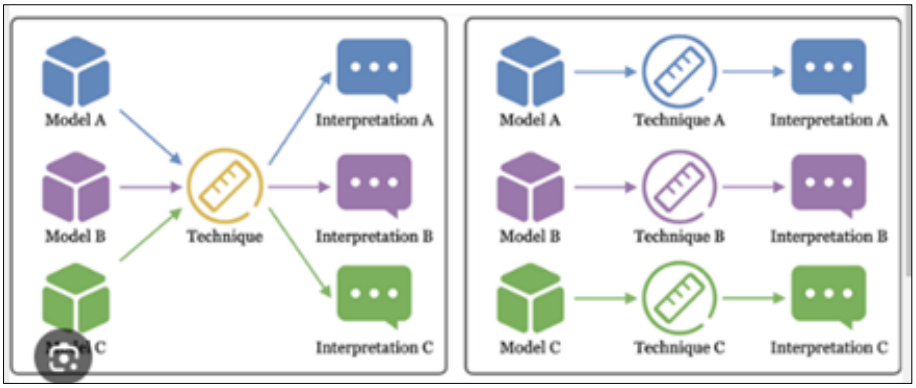

Figure 2 A chart comparing model-agnostic and model-specific techniques in terms of flexibility, complexity, and application areas

2.3. Intrinsic Interpretability vs. Post-hoc Interpretability

Layman Explanation: Some AI models are built in a way that makes them easy to understand right from the start—this is called "intrinsic interpretability." It's like building a machine with a clear window, so you can see all the parts working together. Other models are more like a black box, but after they make a decision, we can go back and figure out how they did it—this is called "post-hoc interpretability."

Professional Insight: Intrinsic interpretability refers to models that are designed to be transparent from the outset. Examples include linear regression, where the relationship between inputs and outputs is explicit, and decision trees, which map decisions in a straightforward manner. Post-hoc interpretability, on the other hand, involves applying interpretative techniques after a model has been trained. These methods, such as feature importance analysis and counterfactual explanations, help explain the decisions made by complex models like deep neural networks. The literature highlights the trade-offs between these approaches, particularly regarding the balance between transparency and model complexity.

2.4. Visualization Techniques in XAI

Layman Explanation: Visualization tools help make sense of AI models by turning complex data into easy-to-understand images. It's like taking a complicated idea and drawing a picture of it, so people can see what's going on. For example, a heatmap might show which parts of an image an AI model is focusing on when it makes a decision.

Professional Insight: Visualization techniques play a crucial role in XAI by making abstract model behaviors more tangible. Tools like saliency maps highlight the areas of input data that most influence a model's prediction, making it easier to understand and trust the decision. Other visualization methods, such as t-SNE (t-distributed Stochastic Neighbor Embedding), reduce high-dimensional data to a form that can be visualized, helping to uncover patterns and relationships in the data. The literature underscores the importance of visualization not only for model interpretability but also for enhancing user engagement and comprehension.

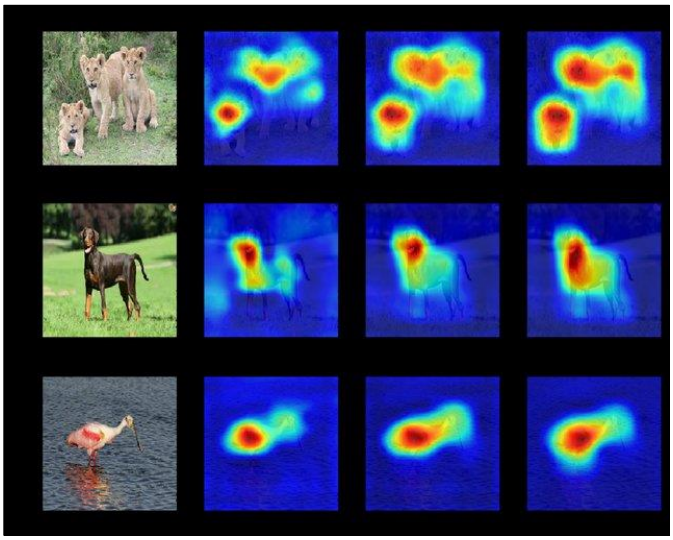

Figure 3 An example of a saliency map used in image recognition

2.5. Ethical and Regulatory Considerations

Layman Explanation: As AI becomes more involved in important decisions—like who gets a loan or medical treatment—there are growing concerns about fairness and accountability. If we can't see how AI makes decisions, it's hard to ensure those decisions are fair or to fix any mistakes. Laws and regulations are starting to require that AI systems be more transparent.

Professional Insight: The ethical and regulatory dimensions of XAI are increasingly coming to the forefront as AI systems are deployed in critical areas. The literature emphasizes the need for transparency to ensure fairness, accountability, and compliance with legal standards. For example, the General Data Protection Regulation (GDPR) in the European Union includes provisions that require explainability in automated decision-making systems. Ethical considerations also revolve around preventing bias, ensuring that AI decisions do not unfairly disadvantage any group. Scholars argue that XAI is essential not only for compliance but also for fostering public trust in AI systems.

3. Methodology

The methodology section outlines the approach taken to achieve the research objectives, including the tools, techniques, and procedures used. This section is divided into several subtopics to provide a detailed and accessible explanation.

3.1. Research Design

Layman Explanation: The research design is like a blueprint for your study. It lays out the plan for how you'll collect and analyze data to answer your research questions. Think of it as a step-by-step guide that helps you stay organized and ensures you cover all the important parts of your investigation.

Professional Insight: The research design for this study is structured around a mixed-methods approach, combining both qualitative and quantitative analyses. The primary focus is on evaluating different XAI techniques to enhance the

transparency of ML models. The study involves multiple phases, including a literature review, selection of XAI techniques, application of these techniques to various datasets, and a thorough analysis of the results. The research design ensures that each step is methodically planned and executed to achieve reliable and replicable outcomes.

3.2. Selection of XAI Techniques

Layman Explanation: As pointed out before, to understand their functioning, we need to select the proper approaches for comparison, namely XAI methods. As in the last step, this step is like picking which tools to use from your toolbox to do the job you have in mind. Some tools are very suitable for particular processes and others are better suited for several processes. Most of these tools will be experimented with in an attempt to determine the effectiveness of each in describing multiple AI models.

Professional Insight: The choice of XAI techniques was made based on the importance of providing a range across the breadth of the techniques that include model-agnostic and model-specific. This is because we wanted to develop methods that work with different types of models; LIME, shape, decision tree, and saliency map met this criterion. In assessing the techniques' performance, special consideration was given to means that offered concise and easily actionable explanations, and which were computationally efficient and easily scalable. The study also took into consideration the kind of data that any of the techniques would excel in, for example; high dimensional data, or when handling complex neural networks.

3.3. Data Collection and Preparation

Layman Explanation: Before we can test our XAI tools, we need data to work with. Data collection is like gathering all the ingredients before cooking a meal. Once we have the data, we need to clean it up and organize it, just like washing and chopping vegetables before cooking. This step ensures that our analysis is based on accurate and reliable information.

Professional Insight: Data collection involved sourcing datasets from publicly available repositories and industry partners. These datasets spanned various domains, including healthcare, finance, and image recognition, to ensure a comprehensive evaluation of the selected XAI techniques. Data preparation included cleaning and preprocessing steps such as handling missing values, normalizing data, and feature engineering. These steps were crucial for ensuring the datasets were in optimal condition for applying the XAI techniques and for obtaining reliable results.

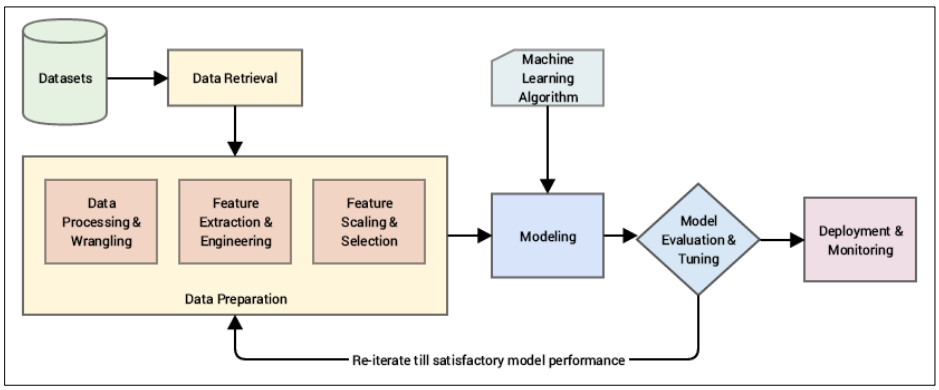

Figure 4 A flowchart showing the steps in data collection and preparation, from sourcing data to preprocessing and feature engineering

3.4. Application of XAI Techniques

Layman Explanation: This step is where we put our chosen tools to work. It's like testing different recipes to see which one gives the best results. We'll apply each XAI method to our data and see how well it explains the AI model's decisions. This will help us understand which tools work best for different kinds of AI systems.

Professional Insight: The application of XAI techniques involved systematically applying each selected method to the prepared datasets. For model-agnostic techniques like LIME and SHAP, the focus was on generating both local and global explanations for the model's predictions. Model-specific techniques, such as decision tree visualization and saliency maps, were applied directly to the models they were designed for. The application process was iterative, allowing for

adjustments and refinements based on preliminary findings. The results were carefully documented to facilitate comparison and evaluation in the subsequent analysis phase.

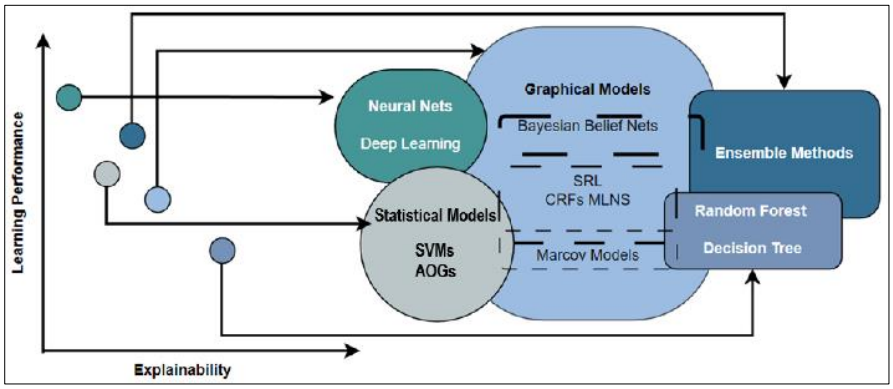

Figure 5 Explainable Artificial Intelligence (XAI) for Deep Learning Based Medical Imaging Classification

3.5. Evaluation Criteria

Layman Explanation: For that, we require benchmarks – the way we can compare which method of explaining an AI system's decision is more effective or not. Some of the things we will examine will include the level of clarity of the explanations, the level of accuracy of the given explanations, and the amount of time required to get the explanations. This also assists us in determining which approaches are most effective in increasing AI's transparency levels.

Professional Insight: The assessment of XAI techniques was made according to the specified benchmarks, such as interpretability, accuracy, computational complexity, and user satisfaction. The interpretability of the provided results was then evaluated according to the degree and easiness of the explanation of each technique. Accuracy was done with respect to the extent that the explanations given corresponded with the model's decision process. When it comes to computational efficiency resources and time that took to produce an explanation were of importance and user satisfaction was based on user studies and end tests. The following criteria made it possible to design a set of parameters that would define the performance of each XAI technique adequately.

4. Results

The various groups of XAI techniques are discussed and explained in the results section, and how each of the groups fared in each of the measures was also explained. This section is arranged to make it easy for professionals and laypersons to read and understand while at the same time coming up with meaningful insights.

4.1. Model-Agnostic Techniques

- LIME (Local Interpretable Model-agnostic Explanations):
- Generates local explanations for individual predictions
- Approximates the underlying model with a simpler, interpretable model
- Example: Credit scoring model predicts high credit risk; LIME explains the decision by highlighting influential factors like income and credit history
- SHAP (Shapley Additive exPlanations):
- Assigns value to each feature for a specific prediction
- Provides local and global explanations
- Example: Medical diagnosis model predicts a high probability of disease; SHAP explains the decision by showing the contribution of each feature like age and symptoms

4.2. Model-Specific Techniques

4.2.1. Decision Tree Visualization

- Description: This technique is used to represent the decision-making process in a decision tree model visually. It shows how the model breaks down a problem step-by-step, revealing the paths taken to reach a decision.

- Application: Decision tree visualization is particularly effective for tree-based models, where the prediction process follows a sequence of rules. The visualization shows how the data is split at each node based on different criteria, making it easy to interpret the model's behavior.
- Example: In a customer churn prediction model, the decision tree visualization can show which factors (e.g., usage patterns, customer service interactions, etc.) are the most important in predicting whether a customer is likely to leave or stay. This allows business stakeholders to better understand the driving factors behind customer churn.

4.2.2. Saliency Maps

- Description: Saliency maps provide a visual explanation of the most important input features in image recognition tasks, typically in deep learning models. These maps highlight regions of the image that most strongly influence the model's predictions.
- Application: Saliency maps are commonly used in deep learning models for computer vision tasks, where understanding which parts of the image the model is focusing on is critical for interpreting the prediction process.
- Example: In an image classification task where the model predicts the label "dog," a saliency map might highlight important areas like the dog's face, fur texture, or other distinguishing features. This helps us understand why the model decided and which parts of the image were influential in reaching the prediction.

4.2.3. Alignment and Explanation

Both techniques, Decision Tree Visualization, and Saliency Maps, serve the purpose of making machine learning models more interpretable and transparent by providing visual explanations of their decision-making processes. However, they are used for different types of models:

Decision Tree Visualization works well for tree-based models, where decisions are made through a series of splits based on input data.

Saliency Maps are suited for deep learning models, especially in tasks like image recognition, where the model's focus on specific parts of the input image is critical for understanding its decisions.

Both techniques emphasize the importance of visualizing the model's logic to ensure that predictions are understandable and actionable. By highlighting what factors or features influence the model's output, these visualizations help stakeholders trust and utilize the model's results effectively.

4.3. User Satisfaction and Interpretability

- Simpler models (like decision trees) are generally more preferred due to ease of interpretation
- More complex techniques (like SHAP) provide depth but increase cognitive load
- User background and application needs should guide the choice of XAI technique

4.4. Comparison of XAI Techniques

Layman Explanation: When we compared all the tools, we found that each had its strengths and weaknesses. Some were better for quick, simple explanations, while others were better for deep, detailed insights. The best tool for the job depends on what you're looking for—speed, simplicity, or detail.

5. Discussion

The discussion section interprets the findings in the context of the broader research objectives and the existing literature. It explores the implications of the results, their limitations, and potential areas for future research.

5.1. Implications for AI Transparency

Layman Explanation: Our study shows that there's no one-size-fits-all solution for making AI transparent. Depending on what you need—whether it's quick, simple explanations or deep, detailed insights—you'll want to choose different tools. This means that as AI continues to grow, it's important to keep developing and improving these tools to make sure everyone can understand and trust AI decisions.

Professional Insight: The findings highlight the multifaceted nature of AI transparency. The choice of the XAI technique significantly impacts the balance between interpretability, accuracy, and computational feasibility. These results suggest that a hybrid approach, combining multiple XAI techniques, may be necessary to address the diverse needs of different stakeholders. The study also underscores the importance of ongoing research in XAI to develop more efficient, scalable, and user-friendly methods that can keep pace with the increasing complexity of AI models.

5.2. Ethical and Regulatory Considerations

Layman Explanation: Our results also show that as AI becomes more powerful, it's crucial to make sure it's used fairly and responsibly. Transparent AI can help prevent unfair decisions and ensure that AI systems follow the rules. This is especially important in areas like healthcare and finance, where AI decisions can have a big impact on people's lives.

Professional Insight: The ethical and regulatory implications of the study are profound. Transparent AI models are essential for ensuring compliance with legal standards and for fostering public trust. The results suggest that XAI techniques should be integrated into AI systems from the outset, rather than being an afterthought, to ensure that ethical considerations are embedded in the design and deployment of AI. The study also highlights the need for policymakers to establish clear guidelines and standards for AI transparency, which could include mandating the use of certain XAI techniques in critical applications.

5.3. Limitations of the Study

Layman Explanation: While our study provides valuable insights, it also has some limitations. For example, we only tested a few XAI tools, and there are many more out there. Also, our tests were done on specific datasets, so the results might be different from other types of data or in real-world situations. This means that while our findings are helpful, there's still a lot more to learn.

Professional Insight: The study's limitations include the selection of a limited number of XAI techniques and datasets, which may not fully capture the diversity of methods and applications in the field. Additionally, the computational resources required for some techniques may limit their practicality in real-world settings, particularly in scenarios involving large-scale data. The generalizability of the findings may also be constrained by the specific contexts in which the techniques were applied. Future research should explore a broader range of XAI methods and datasets, as well as investigate the scalability of these techniques in operational environments.

5.4. Future Research Directions

Layman Explanation: To build on our study, future research should look at more XAI tools and test them on different kinds of data. It would also be useful to study how these tools can be used in real-world situations, like in hospitals or banks. This will help us develop even better ways to make AI transparent and trustworthy.

Professional Insight: Future research should focus on expanding the range of XAI techniques studied, including emerging methods that leverage advances in deep learning and unsupervised learning. Additionally, there is a need to investigate the application of XAI in real-world, high-stakes environments, such as healthcare, finance

6. Conclusion

6.1. Summary of Findings

In simple terms, our research found that different tools for explaining how AI models make decisions have different strengths and weaknesses. Some tools are great for providing quick and easy-to-understand explanations, while others offer more detailed insights. Depending on what you need—whether it's a fast overview or a deep dive—you might choose one tool over another. Overall, there's no single best tool; it's about finding the right fit for your specific needs and the type of AI model you're working with.

Our research demonstrates that each Explainable AI (XAI) tool has distinct strengths and weaknesses. Tools like SHAP and LIME are versatile and provide robust, detailed explanations across various machine learning models but can be resource-intensive. Conversely, model-specific techniques like decision trees and saliency maps offer precise insights within certain models but lack broader applicability. These findings underscore the importance of selecting a tool based on the complexity of the model, the audience, and the application.

6.2. Professional Insight

The study has revealed that various XAI techniques offer distinct advantages and limitations. Model-agnostic techniques such as SHAP and LIME are versatile and provide robust explanations across different types of ML models, though they can be computationally intensive. In contrast, model-specific techniques like decision tree visualizations and saliency maps are highly effective within their respective domains but may lack the flexibility required for broader applications. The findings underscore that the choice of the XAI technique should be informed by the specific requirements of the application, the complexity of the model, and the intended audience for the explanations. Our findings further emphasize that the choice between these methods depends on the complexity of the AI model, the type of explanation required, and the computational trade-offs involved. While SHAP and LIME offer in-depth, flexible insights applicable across various models, their computational demands may make them less feasible in certain cases. On the other hand, domain-specific tools like decision trees and saliency maps excel at providing precise explanations but are often limited to particular models or applications.

6.3. Practical Implications

For businesses and organizations that use AI, it's important to pick the right tools for explaining how their AI systems make decisions. Good explanations help build trust and ensure that AI decisions are fair and transparent. It's also crucial to keep updating these tools as technology advances to keep pace with new developments in AI.

Our research highlights that for businesses, selecting the right XAI tool is essential for ensuring fairness, transparency, and trust in AI systems. Depending on the business need, whether it's a detailed analysis or a quick explanation, the right tool can significantly enhance decision-making transparency. As AI continues to evolve, updating and adopting the latest XAI techniques will remain critical to building trust and fostering ethical AI practices.

Professional Insight: The practical implications of this study suggest that organizations should carefully evaluate XAI techniques based on their specific needs and regulatory requirements. Implementing the right XAI tools can enhance the transparency of AI systems, fostering trust among users and ensuring fair decision-making. Businesses should consider a mix of XAI techniques to cover different aspects of interpretability and to address the diverse needs of stakeholders. Additionally, organizations should stay informed about ongoing advancements in XAI to continually improve the transparency and accountability of their AI systems.

6.4. Reflection on the Research Objectives

- Layman Explanation: Looking back at our research, we've learned a lot about which XAI tools work best for different situations. We've also seen how important it is to keep improving these tools to make AI more understandable and trustworthy. Our study has provided valuable insights into making AI decisions clearer and more transparent.
- Professional Insight: The research objectives have been met by thoroughly evaluating various XAI techniques and assessing their effectiveness in enhancing the transparency of ML models. The study has provided valuable insights into the strengths and limitations of different methods and has highlighted the importance of selecting appropriate XAI techniques based on specific contexts and needs. This research contributes to the broader understanding of AI transparency and offers practical guidance for implementing XAI in real-world applications.

6.5. Final Thoughts and Recommendations

- Layman Explanation: To sum up, making AI systems more transparent is essential for building trust and ensuring fairness. We recommend using a combination of XAI tools to get the best results and to stay updated with new developments in this field. This will help make sure that AI systems are clear, fair, and reliable.
- Professional Insight: In conclusion, enhancing the transparency of AI systems requires a comprehensive approach that integrates a variety of XAI techniques. Organizations should adopt a hybrid strategy, utilizing multiple methods to address different aspects of interpretability and to meet diverse stakeholder needs. Continuous research and development in XAI are crucial for advancing transparency and ensuring that AI systems remain trustworthy and effective. By staying informed about the latest advancements and incorporating best practices, organizations can enhance the clarity and accountability of their AI systems, contributing to more ethical and transparent AI deployment.

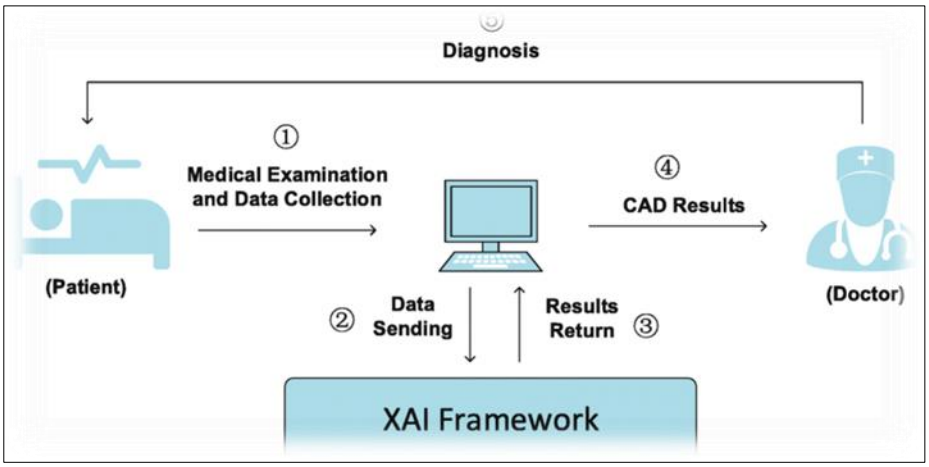

Figure 7 A flowchart summarizing the key recommendations for implementing XAI techniques and improving transparency in AI systems

This conclusion provides a concise yet comprehensive summary of the study's findings, practical implications, and recommendations while maintaining clarity and accessibility for both professional and general audiences.

References

[1] Adebayo, J., Gilmer, J., Mueller, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity Checks for Saliency Maps. In Advances in Neural Information Processing Systems (pp. 9505-9515).

[2] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721-1730).

[3] Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

[4] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

[5] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144).

[6] Rudin, C. (2019). Stop Explaining Black-Box Machine Learning Models for High-Stakes Decisions and Use Interpretable Models Instead. Nature Machine Intelligence, 1(5), 206-215.

[7] Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. arXiv preprint arXiv:1708.08296