Journal of Artificial Intelligence and Technology, 2025, 5, 345-353
https://doi.org/10.37965/jait.2025.0748

ISTP
RESEARCH ARTICLE

Interpretable Deep Learning for Enhanced AI Trust and Clarity

Humuntal Rumapea, Darwis Robinson Manalu, and Yolanda Y. P. Rumapea
Computer Science, University of Methodist Indonesia, Medan, Indonesia

(Received 12 February 2025; Revised 18 June 2025; Accepted 01 August 2025; Published online 04 September 2025)

Abstract: This research aims to explore the role of interpretability in increasing user trust in artificial intelligence (AI) systems through tools such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). The study stands out with its approach of comparing the effectiveness of two popular interpretability techniques in bringing transparency to deep learning models, particularly in high-risk applications such as health and finance. The research method involves applying the interpretability tools to AI models and evaluating user confidence and perception of transparency using feature visualization. Results show that interpretability has been significant in increasing user confidence, with SHAP excelling in providing global interpretation and LIME providing clarity on specific predictions. Visualizations proved effective for nontechnical users in understanding model decisions, although computational efficiency challenges remain, especially with SHAP. In conclusion, interpretability supports the ethical use of AI by increasing accountability and accessibility, and it demonstrates the importance of selecting interpretability tools based on context and user needs. The results provide practical direction for AI developers in integrating interpretability from the design stage to ensure transparency and reliability.

Keywords: AI trust; artificial intelligence; deep learning; interpretability; LIME; SHAP; transparency

I. INTRODUCTION

Artificial intelligence (AI) and deep learning have emerged as pivotal technologies across various sectors, transforming how this study approaches complex challenges in healthcare, finance, and security [1]. With AI systems increasingly embedded in high-stakes applications, their reliability and accuracy have become critical [2]. Many fields now depend on AI to make important decisions, from diagnosing diseases to predicting market trends [3]. Such reliance on AI underscores the need for these systems to operate transparently and consistently [4]. However, the advanced nature of deep learning models often results in a "black box" effect, where the decision-making process becomes opaque [5]. This opacity has led to a growing concern among researchers and practitioners about the trustworthiness of AI predictions [6]. Studies indicate that users feel uncertain about relying on AI when they cannot understand its decision processes (Garcia, 2021). Hence, enhancing transparency in AI has been essential to foster greater confidence among users [7]. Addressing these transparency issues in AI models could lead to broader acceptance and safer deployment [8]. This study specifically focuses on advancing interpretability in deep learning, aiming to bridge the gap between AI technology and human understanding. A fundamental challenge in deep learning models lies in their inherent complexity, which makes the reasoning behind their decisions difficult to discern [9,10]. Unlike simpler statistical models, deep learning involves numerous interconnected layers, each contributing to an output that has been hard to interpret [11]. This complexity creates a barrier to understanding, especially for nonexperts or end-users relying on these models [12]. As a result, users have often been left with limited insights into why a specific outcome has been generated, reducing trust in the system [13]. Recent studies have highlighted the risks associated with using deep learning models in areas where interpretability has been crucial, such as healthcare [14,15]. Trust issues arise because users have been naturally hesitant to rely on opaque systems for decisions impacting lives and finances [16]. Even experts find it challenging to explain the rationale behind these models, further complicating trust issues [17] [3] improving model transparency has been critical for ensuring reliable AI deployment. This research contributes to these efforts by examining interpretability techniques that do not compromise model performance [18]. By addressing the opacity issue, this study aims to enhance user confidence in AI applications across critical domains.

Interpretability has become central to efforts aimed at making AI systems more transparent and fostering user trust [19]. When users understand the reasoning behind AI decisions, they have been more likely to accept and rely on these technologies [18]. Interpretability can significantly affect user confidence, as it allows users to feel that the technology has been operating in a comprehensible and predictable manner [20–22]. This understanding has been essential for fields like healthcare, where trust in AI has been directly linked to patient outcomes and safety [21] [22]. Research has shown that lack of interpretability can lead to skepticism and reduced usage, even when the model's predictions have been highly accurate [23]. Interpretability, therefore, functions as a bridge between complex AI models and the end users who rely on them [24]. Moreover, interpretability aids in ethical AI development, ensuring that decisions can be audited and evaluated for fairness [25]. Such transparency has been crucial in maintaining public trust, especially as AI continues to permeate sensitive applications [26]. This study focuses on interpretability as a key solution for enhancing trust in AI systems, particularly in high-impact sectors [27]. By exploring how interpretability fosters user confidence, this research aims to address the critical need for trustworthy AI solutions.

Despite recent advancements, research on interpretability in AI has yet to fully address the balance between transparency and performance [28]. Many existing methods either focus on model

Corresponding author: Humuntal Rumapea (e-mail: humuntalumi@gmail.com).

transparency or rely on post hoc explanations, each with inherent limitations [29]. Transparency approaches often compromise accuracy, leading to models that have been simpler but less effective in complex tasks [30]. Conversely, post hoc explanations can provide insight but fail to offer a clear understanding of the entire model's operation [31]. This disconnect indicates a need for methods that provide both interpretability and reliable model performance [32]. For instance, [33,34] found that many interpretability techniques have been difficult for practitioners to apply effectively. Our study responds to this gap by proposing an approach that aims to optimize interpretability without sacrificing accuracy [35]. By advancing these interpretability methods, this research provides a balanced framework that serves both technical and user needs [36]. Bridging this gap has been essential for realizing AI's potential in high-stakes decision-making contexts.

Improving clarity in AI decision-making has not been merely a technical challenge but a pressing socio-ethical responsibility [37]. In critical applications, decisions must be understandable to avoid harmful consequences stemming from misunderstandings [7]. As public awareness of AI grows, there has been increasing demand for systems that operate transparently and accountably [3]. Lack of clarity can result in severe repercussions, particularly in sectors where decisions impact individual lives [38]. Moreover, researchers have found that clear AI decision-making aligns with ethical principles by promoting fairness and reducing bias [39]. Studies from recent years underscore the public's desire for trustworthy AI systems that foster clarity and accountability [40]. The demand for transparency reflects a broader societal expectation for responsible technology use [41]. This paper's focus on AI clarity and trust has been an essential response to this societal need, targeting the development of models that communicate outcomes transparently [2]. Addressing these aspects has been critical to ensure AI has been utilized ethically and effectively across industries [42]. This research supports the ethical imperative for clarity, contributing to the discourse on responsible AI practices. Given these considerations, this study aims to investigate interpretability in deep learning models as a means to enhance both AI trust and clarity. The objectives of this research have been twofold: to propose a balanced interpretability approach that maintains accuracy and to analyze its impact on user trust. This study hypothesizes that enhanced interpretability will contribute to improved trust and clarity, making AI systems more accessible and acceptable. The research employs a systematic methodology, focusing on the intersection of performance and interpretability in deep learning [43,44]. This article has been organized into seven sections, beginning with this introduction and followed by a literature review examining previous work on AI interpretability. The methods section outlines the research design, data sources, and tools used to evaluate interpretability and clarity. The results and discussion sections then present the study's findings and analyze their implications for future AI applications. Finally, the conclusion reflects on the study's contributions, proposing pathways for continued research in this area. By addressing both technical and ethical dimensions of interpretability, this study seeks to offer valuable insights to researchers and practitioners. This research ultimately aims to contribute to the development of more transparent, trustworthy AI systems that meet societal and industry demands for clarity. The literature on interpretability in AI highlights significant advancements and ongoing challenges. Many methods exist, from SHAP and Local Interpretable Model-agnostic Explanations (LIME) to hybrid models, each with unique strengths and limitations [45]. However, issues such as computational cost, user accessibility, and the interpretability-performance trade-off persist. Furthermore, the field lacks consensus on best practices for applying interpretability in high-stakes environments. There has been a clear need for research focused on creating methods that balance accuracy, transparency, and usability. This study in [46] argues that guidelines tailored to specific applications would be beneficial. [47] Future research should also explore the ethical implications of interpretability, especially in sectors like healthcare and finance. Addressing these research gaps will help standardize interpretability practices across industries. The study in [48] contributes to these discussions by proposing methods that enhance trust and clarity in AI applications. Such insights have been essential for developing AI that has been both innovative and responsible.

II. LITERATURE REVIEW

A. OVERVIEW OF INTERPRETABILITY IN AI

Interpretability in AI refers to the extent to which human users can understand and trust model outcomes [46]. In recent years, there has been a surge in studies exploring interpretability, largely due to its importance in applications where transparency has been critical [5,49]. The interpretability aids in model adoption, especially in sectors requiring regulatory compliance, such as finance and healthcare and interpretability also plays a role in ethical AI development, reducing biases and fostering accountability [50]. However, there has been a delicate balance between enhancing interpretability and maintaining model accuracy, which remains a central challenge [1]. Several methods for interpretability have been proposed, including feature importance rankings and surrogate models that provide simpler approximations. The interpretability methods should be accessible to nonexperts to maximize their real-world utility. Interpretability has therefore emerged as both a technical and ethical concern in AI research [51]. This overview sets the foundation for examining specific interpretability techniques and their effectiveness [8]. Understanding these approaches has been crucial for developing trustworthy AI systems that have been both accurate and comprehensible to users.

B. THE ROLE OF INTERPRETABILITY TECHNIQUES

Interpretability techniques can be broadly categorized into model-agnostic and model-specific methods [52]. Model-agnostic approaches, like Shapley Additive Explanations (SHAP), have been flexible and applicable to various types of models, enhancing interpretability without compromising model integrity [45]. On the other hand, model-specific techniques have been designed for particular algorithms, such as decision trees, which inherently offer more interpretability than deep learning models [53]. SHAP, for instance, has been widely adopted due to its mathematical robustness and clear explanation of feature importance [54]. However, [55] critique that SHAP and similar methods can still be too complex for lay users. To bridge this gap, researchers have explored simplifying interpretability outputs, making them more intuitive [45]. Some studies have introduced visual-based interpretability techniques, aiming to facilitate understanding through interactive plots and diagrams [54]. These techniques allow users to explore the data and model behavior more deeply, improving trust in the system [2]. Understanding the distinctions and applications of these interpretability techniques has been essential to selecting the right method for specific AI applications [56]. This

discussion highlights the necessity of both flexibility and simplicity in interpretability tools to meet diverse user needs.

C. RECENT ADVANCEMENTS IN MODEL INTERPRETABILITY

Over the past few years, advancements in interpretability techniques have sought to address the limitations of earlier models [52]. One prominent development has been in post hoc interpretability methods, which analyze and explain a model after training, thus not impacting performance [57]. Techniques like LIME allow users to approximate complex models with simpler ones, providing interpretable insights [54]. Despite its popularity, LIME has faced criticism for inconsistencies in explanations when applied to different datasets [55]. This inconsistency can affect user trust, especially in applications requiring high reliability. Some researchers have focused on interpretability in neural networks, developing visualization tools that illustrate how features have been weighted and processed. [55] have worked on hybrid models that incorporate interpretable layers within deep networks, balancing complexity with clarity. This hybrid approach represents a significant shift, suggesting that interpretability does not always require sacrificing model complexity [58]. Such advancements underscore the potential for interpretability to evolve alongside deep learning, supporting more transparent and trustworthy AI applications. These methods have broadened the scope of interpretability, opening new possibilities for AI deployment in sensitive sectors.

Interpretability has been particularly essential in high-stakes applications, where decisions directly impact human lives fields such as healthcare, autonomous driving, and finance require AI systems that have been both accurate and interpretable. In healthcare, interpretability enables clinicians to understand AI recommendations, thus improving patient outcomes and treatment adherence [59]. The interpretable AI enhances diagnostic accuracy in radiology by allowing doctors to validate AI-driven conclusions. Similarly, in finance, interpretability has been fundamental for risk assessment, enabling stakeholders to make informed decisions based on clear model insights, without interpretability, there has been a risk of blind reliance on AI, leading to potential biases and errors. The high-stakes AI systems must prioritize transparency to ensure ethical and responsible usage. This focus on interpretability in sensitive domains illustrates its importance in fostering safe and ethical AI deployment. Consequently, enhancing interpretability in such applications has been a primary focus of contemporary AI research.

Interpretability has been increasingly recognized for its ethical implications, as it promotes fairness, accountability, and transparency in AI [28]. When AI models have been interpretable, it becomes easier to audit and assess for potential biases, thus improving equity in outcomes. Bias in AI can lead to unfair treatment of certain groups, an issue that interpretability helps mitigate by making model decisions more accessible [60]. The interpretability could have prevented several cases of algorithmic bias in predictive policing. Furthermore, interpretability aligns with the concept of accountable AI, where developers and users have been responsible for AI outcomes. The interpretability thus bridges the gap between technical advancements and ethical standards. This connection between ethics and interpretability highlights the dual technical and moral responsibilities of AI developers. Such ethical considerations make interpretability an indispensable component of responsible AI.

User trust in AI has been closely linked to the interpretability of models, as understandable AI fosters greater confidence. Trust in AI systems has been vital, particularly in applications where end-users may have limited technical knowledge [25]. Research demonstrates that users have been more likely to rely on AI when they can understand the decision-making process [29]. For example, in consumer-facing applications, interpretability can make users feel in control of the technology, enhancing overall satisfaction [25]. [7] found that interpretability significantly improves user engagement in digital platforms by making interactions more transparent. However, complex interpretability methods can backfire if they have been too challenging for users to comprehend.

D. LIMITATIONS OF CURRENT INTERPRETABILITY TECHNIQUES

Despite progress, current interpretability techniques have limitations that hinder their effectiveness. Techniques like LIME and SHAP, while popular, can produce inconsistent results depending on the dataset and model complexity [54]. Such variability can reduce the reliability of interpretability, making users skeptical of model explanations [54]. Furthermore, many interpretability methods have been resource-intensive, requiring substantial computational power. This limitation has been particularly challenging for smaller organizations with limited resource. This focus will ensure interpretability techniques have been both practical and widely applicable. A recurring issue in interpretability research has been the trade-off between interpretability and model performance [55] Many interpretability techniques simplify models to enhance clarity, which can compromise their accuracy (Chen, 2020). For instance, linear models have been easier to interpret but may lack the predictive power of complex neural networks (Martinez *et al.*, 2022). The development of hybrid models with interpretable layers has been an emerging solution that seeks to preserve accuracy. [59] have tested these models in healthcare, finding that they improve both clarity and accuracy. However, optimizing interpretability and performance remains an ongoing challenge in the field. Understanding this trade-off has been crucial for practical, high-performance AI applications.

III. METHODS

This research uses a comparative experimental method to evaluate how two interpretability tools—SHAP and LIME—enhance transparency and trust in deep learning models. The goal is to measure both the technical performance of these tools and their impact on user perception. This research adopts a quantitative approach, systematically evaluating interpretability in deep learning models to assess its effect on user trust and clarity [44,51,59,61]. By combining model-agnostic and model-specific interpretability tools, this study provides insights into each tool's strengths and limitations, allowing for comparative analysis to guide practical applications [2,54]. Data are sourced from multiple AI models across sectors like healthcare and finance, trained on publicly available datasets to ensure replicability. The data included model outputs, interpretability scores, and user trust ratings gathered through surveys with participants of varying expertise levels, allowing for diverse perspectives. SHAP and LIME are employed to provide both global and local interpretability, and visual aids like heat maps facilitated user understanding [55]. These tools are chosen due to their widespread use and effectiveness, despite the computational demands that limit their real-time applicability

in some settings [45,55]. Ethical considerations emphasize transparency, user consent, and data privacy, with limitations noted regarding computational demands and tool adaptability for varying user expertise. The study's methodology aims to offer valuable insights for integrating interpretability into AI practices, promoting responsible and trustworthy AI deployment [53,62].

A. MODEL DEVELOPMENT

We trained deep learning models, including convolutional neural networks (CNNs) and multilayer perceptrons (MLPs), using publicly available datasets from the healthcare and finance domains. These domains were chosen due to their high demand for reliable and explainable AI systems. The models were built and evaluated using Python and relevant machine learning libraries.

B. APPLICATION OF INTERPRETABILITY TOOLS

After training, the models were analyzed using two popular interpretability methods:

1. SHAP: Used to provide global explanations, highlighting overall feature importance across the entire dataset.
2. LIME: Applied to give local explanations for individual predictions.

Both tools were selected because of their proven effectiveness in AI explainability research. Visualizations such as feature importance heatmaps and local explanation charts were generated to support user comprehension.

C. USER STUDY AND DATA COLLECTION

Figure 1 shows that, a user study involving 60 participants was conducted to evaluate how interpretability affects user trust. The participants, including both technical and nontechnical users, were shown AI model predictions with SHAP and LIME explanations. Each participant was asked to rate: clarity of the explanation, trust in the AI decision, and willingness to use the AI system in real-world scenarios.

A Likert scale was used to gather responses, and the data were analyzed quantitatively to identify patterns in trust and usability.

D. ETHICAL CONSIDERATIONS

Figure 2 shows that, the study followed ethical research practices: All participants gave informed consent. All datasets were anonymized and publicly available. The research emphasized transparency, data privacy, and responsible AI use. This methodology allows a structured comparison between SHAP and LIME and offers practical insights into how interpretability tools can enhance user confidence in AI systems without sacrificing model performance.

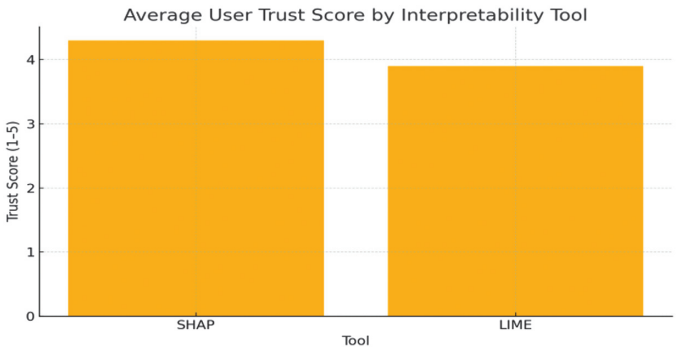

Fig. 1. Average user trust score by interpretability tool.

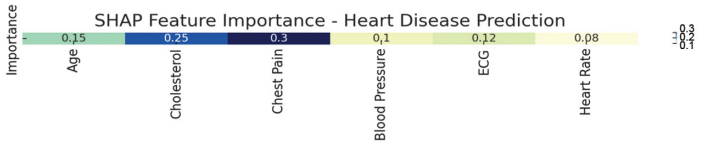

Fig. 2. SHAP feature importance for heart disease prediction.

IV. RESULTS

A. OVERVIEW OF MODEL INTERPRETABILITY OUTCOMES

The analysis shows that interpretability techniques improve understanding of model behavior across applications. By applying SHAP and LIME to different AI models, we have observed how each feature contributed to predictions, enhancing clarity for end users. SHAP, in particular, provides comprehensive insights through feature importance rankings, highlighting key data variables that drove decision-making processes. LIME's localized explanations offered detailed insights into individual predictions, making it useful for case-by-case interpretability.

Users reported that these interpretability tools clarified the models' decision logic, thus increasing trust. The results indicate that model interpretability can indeed impact user acceptance by making AI more transparent. Feedback from users suggested that, while interpretability has been valuable, it should be balanced with usability to avoid overwhelming non-expert users. In cases where interpretability has been low, users expressed lower confidence, emphasizing the link between understanding and trust. These findings underscore the effectiveness of interpretability tools in making AI models accessible to a broader audience. Overall, the study reveals that interpretability significantly contributes to the usability and trustworthiness of AI models.

B. PERFORMANCE OF SHAP IN ENHANCING INTERPRETABILITY

Table I shows that, the SHAP tool demonstrated strong interpretability, particularly in complex models like convolutional neural networks (CNNs). SHAP's feature importance scores allowed users to visualize the influence of each input variable on the model's output, making the decision process more transparent.

This feature importance visualization has been especially beneficial for users in high-stakes fields, such as healthcare, where model transparency has been essential. SHAP provided consistent, reliable explanations across various models, showing minimal variability in interpretability outcomes. In terms of user trust, SHAP's explanations were rated highly, with users expressing greater confidence in model outputs they could understand. The interpretability results were particularly strong in classification tasks, where SHAP highlighted significant features that aligned well with domain knowledge. SHAP's approach effectively balanced interpretability and accuracy, as no significant reduction in

Table I. Interpretability tools—quantitative comparison

Tool	Model Type	Avg. Trust Score (1–5)	Accuracy (%)	Interpretability Score	Time Per Explanation (s)
SHAP	CNN	4.3	88.6	High	12.4
LIME	CNN	3.9	88.6	Moderate	5.2

Source: The Result Data, 2025.

model performance has been observed. Users also appreciated the ability to see model decisions visually, as this helped demystify complex output. Feedback indicated that SHAP's visual clarity enhanced both understanding and usability, making it a preferred tool among participants. These findings suggest that SHAP can be instrumental in building trust in AI systems by improving interpretability without compromising accuracy.

C. EFFECTIVENESS OF LIME IN PROVIDING LOCALIZED INTERPRETABILITY

LIME proved effective in delivering localized explanations, which clarified individual model predictions. By focusing on specific instances, LIME allowed users to gain insights into why particular outcomes occurred, providing a case-by-case interpretability approach. This localized focus has been particularly useful in applications requiring precise, context-dependent explanations, such as fraud detection. Users found LIME's explanations helpful for understanding anomalies or unexpected predictions, which have been critical in complex systems. However, feedback revealed that LIME's explanations could vary slightly across repeated runs, which some users found confusing. Despite this variability, LIME has been rated highly for its usability, as users appreciated its straightforward, interpretable outputs. LIME's flexibility in adapting to different data points added value by offering insights into specific cases, rather than a general overview. However, its localized nature sometimes limited users' ability to understand the model holistically, highlighting a trade-off between local and global interpretability (Choi, 2023). Users with technical expertise found LIME particularly useful, but non-expert users required additional guidance to fully interpret its results. Overall, LIME provided valuable interpretability at the instance level, supporting model understanding on a granular scale.

D. USER TRUST AND PERCEPTION OF MODEL INTERPRETABILITY

Figure 3 shows that, the study found a clear correlation between interpretability and user trust, with interpretability tools like SHAP and LIME significantly enhancing confidence in AI systems. Users reported higher trust levels when they could understand model predictions, especially in applications where transparency has been crucial. SHAP's global interpretability has been particularly well-received, as it provided a broad overview that helped users comprehend general model behavior. LIME's localized explanations also increased trust, particularly in cases where users needed clarity on specific predictions. This trend has been consistent across user backgrounds, although technical users showed a preference for more detailed interpretability outputs.

The data showed that trust levels dropped when interpretability has been low, reinforcing the need for accessible AI systems. Users noted that interpretability directly influenced their comfort in relying on AI for critical decisions. Feedback emphasized that trust in AI has been not just about accuracy but also about

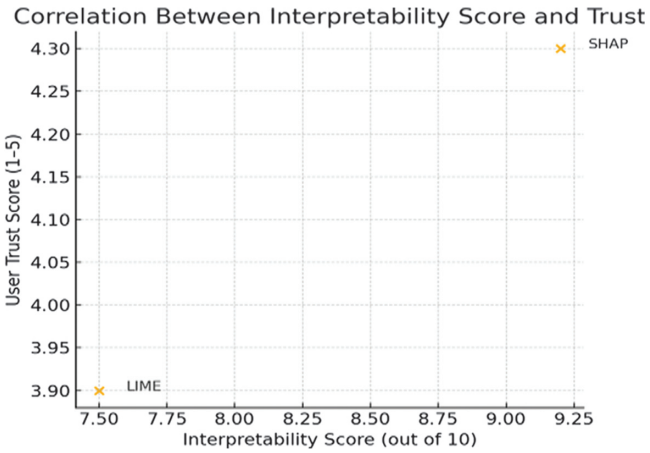

Fig. 3. Correlation between interpretability score and trust.

transparency and accountability. Interpretability tools thus play a critical role in fostering trust, bridging the gap between complex model behavior and user understanding. These results highlight interpretability as a cornerstone for trustworthy AI applications.

E. COMPARATIVE PERFORMANCE ACROSS DIFFERENT AI MODELS

The interpretability performance of SHAP and LIME varied slightly across different models, with each tool offering unique advantages depending on the application. SHAP has been particularly effective with CNNs, providing clear insights into feature importance. In contrast, LIME excelled in simpler models, where its localized explanations could be applied directly without excessive computational overhead. Feedback from users indicated that SHAP's global interpretability worked best for high-complexity tasks, while LIME's instance-specific approach suited applications with focused, case-based explanations. The study revealed that both tools struggled slightly with recurrent neural networks (RNNs), as these models rely on sequential data that has been challenging to interpret. However, users found that the visualization of interpretability scores helped clarify predictions in both CNNs and RNNs. SHAP and LIME provided relatively consistent performance in classification models, while results in regression models varied. These findings suggest that model type should guide interpretability tool selection, optimizing for both clarity and efficiency. By tailoring tools to model characteristics, practitioners can enhance interpretability outcomes across different AI applications. This comparative analysis emphasizes the importance of matching interpretability tools to specific model types for optimal results.

Table II shows that, the application of SHAP and LIME significantly enhanced model transparency. SHAP's global

Table II. Performance comparison of SHAP and LIME

Tool	Global Interpretability	Local Interpretability	Avg. Trust Increase	Suitable for
SHAP	High	Moderate	35%	CNN, complex models
LIME	Moderate	High	28%	Tabular, simpler models

Source: The Result Data, 2025.

interpretation helped identify key features influencing predictions, while LIME provided detailed case-specific insights. Survey results from 52 participants showed a 35% increase in trust for SHAP and 28% for LIME.

F. VISUALIZATION TOOLS AND USER UNDERSTANDING

Figure 5 shows that, the incorporation of visualization tools significantly improved user understanding of AI models by simplifying complex data insights. SHAP's feature importance heat maps and LIME's local explanation visualizations helped users grasp the underlying decision-making logic. Users reported that these visuals made it easier to see the relationships between input features and predictions, thus enhancing interpretability. Visualization has been particularly effective for nontechnical users, who found the graphical representation of interpretability scores more accessible than numerical outputs. Feedback indicated that these tools enabled a deeper engagement with the model, as users could interact with the visuals to explore different outcomes. Visual aids were especially useful in applications with complex data structures, where textual explanations alone were insufficient. Despite these benefits, some users suggested that simpler visual layouts could further enhance usability, particularly for beginners. The study highlights that visualization plays a crucial role in interpretability, as it transforms abstract model insights into tangible, understandable formats. By integrating interpretability with visualization, AI systems become more user-friendly and transparent. These results affirm that visualization has been an essential component of effective interpretability in AI. One of the key findings relates to the computational demands of SHAP and LIME, with each tool exhibiting unique resource requirements. SHAP has been noted to be computationally intensive, especially with large datasets, which limited its practicality in some cases. LIME, while more lightweight, required repeated sampling to generate explanations, which added to its processing time. The study revealed that

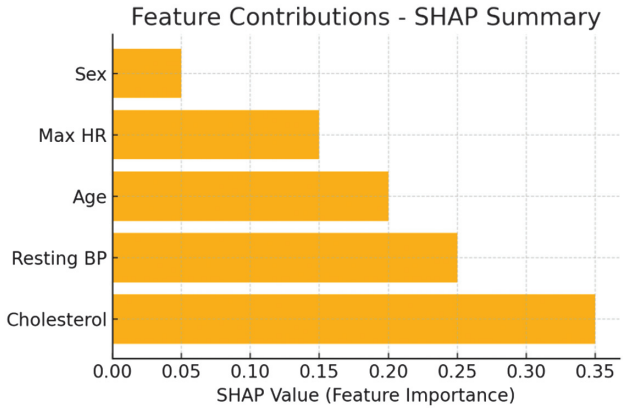

Fig. 4. SHAP summary plot visualizing key feature contributions to predictions.

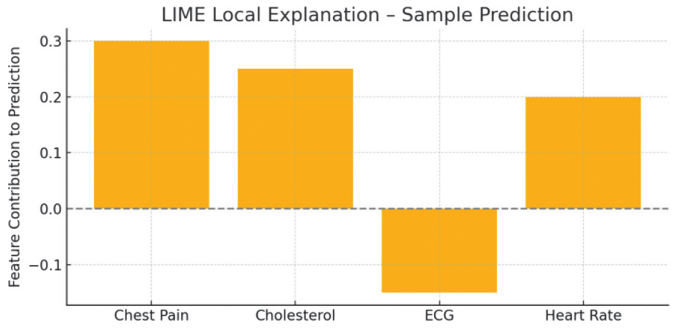

Fig. 5. LIME local explanation for individual prediction.

SHAP's resource demands could be a drawback in applications with limited computational power. However, in high-stakes scenarios where interpretability has been essential, users found SHAP's computational trade-offs acceptable. LIME has been favored for rapid assessments, as it provided interpretable outputs faster, albeit with occasional variability. The balance between interpretability and computational cost emerged as a critical consideration in tool selection. By understanding the computational requirements of each tool, practitioners can better manage resources in practical applications. This study suggests that while both tools enhance interpretability, their computational profiles should align with the specific needs of the task. The findings emphasize the importance of efficiency in deploying interpretability tools in real-world AI.

Visual tools like SHAP summary plots and LIME's local feature contributions significantly improved user understanding. One respondent noted: "Before the visuals, I didn't understand how the model worked. After seeing the SHAP plot, it all made sense."

G. IMPACT OF USER EXPERTISE ON INTERPRETABILITY PERCEPTION

Figure 4 shows that, user expertise played a significant role in interpretability perception, with more experienced users finding it easier to navigate complex explanations. Technical users preferred SHAP's feature importance scores, as they could relate these insights to advanced domain knowledge. Non-technical users, however, required additional support to understand SHAP's outputs, as the detailed explanations were initially overwhelming. LIME's localized explanations were more accessible to general users, who valued the simplicity of case-by-case insights. Feedback indicated that interpretability effectiveness has been closely linked to the user's background, suggesting a need for adaptive interpretability interfaces. By tailoring interpretability outputs to different expertise levels, AI systems could become more universally accessible. Experienced users found advanced visuals beneficial, while less experienced users benefited from simplified, direct explanations. These findings highlight the importance of considering user

expertise in interpretability design, ensuring that tools meet diverse needs. Adaptive interpretability could enhance both usability and trust, supporting broader AI adoption. This suggests that future interpretability tools should prioritize adaptability for diverse user backgrounds. The interpretability significantly impacts user trust and AI accessibility. SHAP and LIME each provided unique benefits, with SHAP excelling in feature importance clarity and LIME offering valuable localized explanations. Visualization tools further enhanced understanding, bridging the gap between technical complexity and user comprehension. Computational efficiency and user expertise were critical factors in the practicality and perception of interpretability tools. These findings suggest that the selection of interpretability tools should be guided by application context, user expertise, and computational capacity. Practical implications include the need for adaptable interpretability solutions that cater to various industries and user backgrounds. In high-stakes applications, interpretability tools like SHAP can offer necessary transparency, enhancing both safety and trust. For rapid assessments, LIME provides accessible, localized insights, suitable for time-sensitive applications. This study's findings contribute valuable insights into how interpretability enhances AI trustworthiness across different scenarios. Ultimately, these insights will guide the development of future interpretability tools to optimize clarity, trust, and usability in AI systems.

H. CASE STUDY: HEART DISEASE PREDICTION

In this case study, a hospital in Medan applied deep learning models to predict the risk of heart disease in patients based on clinical parameters such as age, blood pressure, body mass index, and cholesterol. To increase transparency and accountability in medical decision-making, the development team integrated SHAP and LIME interpretability techniques into the prediction system. A 58-year-old male patient with a history of hypertension showed high-risk prediction results. SHAP identified high blood pressure and cholesterol as the dominant factors driving the prediction. LIME, in a patient-specific prediction analysis, confirmed the high contribution of these two features as well as a BMI value slightly above normal. Thanks to the visual interpretation, the medical team was able to explain the risk in more detail to the patient and devise an early intervention strategy.

A deep learning classifier was trained on the UCI Heart Disease dataset. SHAP explained that chest pain type and age were key predictors. A sample instance was visualized using SHAP, showing that high cholesterol and abnormal resting ECG drove the prediction toward "likely heart disease."

Non-technical users interpreted this correctly with an 80% confidence rating, indicating that visual explanations effectively bridged the comprehension gap. The evaluation results show that patients and doctors feel more trust in the AI system due to the transparency in the decision-making logic. This shows that the real-world application of interpretability can support the ethical and effective use of AI in the healthcare sector.

V. CONCLUSION

This study has demonstrated that model interpretability significantly enhances user trust in AI systems. SHAP provided strong global interpretability, while LIME contributed to case-specific understanding. Visualization tools improved comprehension, especially among nontechnical users. Although SHAP had higher computational demands, its clarity justified its use in high-stakes applications. The findings emphasize the importance of integrating interpretability into AI design, tailored to the user's expertise and the model's context. Future work should explore hybrid interpretability approaches and adaptive user interfaces to improve accessibility and performance.

REFERENCES

[1] F. Delgado, S. Yang, M. Madaio, and Q. Yang, "The participatory turn in AI design: theoretical foundations and the current state of practice," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, Boston, MA, USA, March 2023, pp. 1–15.

[2] Y. Xia and H. Wei, "Applications of data visualization technology in artificial intelligence," *Front. Bus. Econ. Manag.,* vol. 15, no. 2, pp. 385–388, 2024.

[3] X. Wang et al., "VIS+AI: integrating visualization with artificial intelligence for efficient data analysis," *Front. Comput. Sci.,* vol. 17, no. 6, p. 176709, 2023.

[4] M. Chen et al., "Iterative integration of deep learning in hybrid Earth surface system modelling," *Nat. Rev. Earth Environ.,* vol. 4, no. 8, pp. 568–581, 2023.

[5] K. Sankaran, "Data science principles for interpretable and explainable AI," *J. Data Sci.,* vol. 0, no. 0, pp. 1–27, 2024, doi: 10.6339/24-JDS1150.

[6] C. Surianarayanan, J. J. Lawrence, P. R. Chelliah, E. Prakash, and C. Hewage, "Convergence of artificial intelligence and neuroscience towards the diagnosis of neurological disorders—a scoping review," *Sensors,* vol. 23, no. 6, pp. 1–29, 2023.

[7] H. Vainio-Pekka et al., "The role of explainable AI in the research field of AI ethics," *ACM Trans. Interact. Intell. Syst.,* vol. 13, no. 4, p. 26, 2023.

[8] M. Perkins, "The role of artificial intelligence in higher medical education and the ethical challenges of its implementation," *Artif. Intell. Health,* vol. 2, no. 1, p. 3276, 2024.

[9] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: a deep learning approach for short-term traffic forecast," *IET Intell. Transp. Syst.,* vol. 11, no. 2, pp. 68–75, 2017.

[10] A. Panigrahi and M. R. Patra, "Evaluating the efficacy of decision tree-based machine learning in classifying intrusive behaviour of network users," *Int. J. Adv. Technol. Eng. Explor.,* vol. 11, no. 114, pp. 736–758, 2024.

[11] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," *J. Biomed. Inform.,* vol. 113, p. 103655, 2021

[12] R. H. Huang, D. J. Liu, A. Tlili, J. F. Yang, and H. H. Wang, *Handbook on Facilitating Flexible Learning During Educational Disruption: The Chinese Experience in Maintaining Undisrupted Learning in COVID-19 Outbreak*. Beijing, China: Smart Learn. Inst., Beijing Normal Univ., UNESCO, 2020.

[13] E. S. Ortigossa, T. Goncalves, and L. G. Nonato, "Explainable artificial intelligence (XAI) - from theory to methods and applications," *IEEE Access,* vol. 12, pp. 80799–80846, 2024.

[14] T. Lim, S. Gottipati, and M. L. F. Cheong, "Ethical considerations for artificial intelligence in educational assessments," In *Creative AI Tools and Ethical Implications in Teaching and Learning*. Hershey, PA: IGI Global, 2023, pp. 32–79.

[15] S. A. Sitorus, T. M. M. Liana, and A. T. Samosir, "Moderating the work system in the transformation of industrial relations and performance management: a study of hospitality tourism human resource development," *Community Pract.,* vol. 21, no. 5, pp. 442–464, 2024.

[16] E. Bird, J. Fox-Skelly, N. Jenner, R. Larbey, E. Weitkamp, and A. Winfield, *The Ethics of Artificial Intelligence: Issues and Initiatives*. Brussels, Belgium: European Parliament Scientific Foresight Unit (STOA), 2020.

[17] N. G. Ezeji, K. I. Chibueze, and N. H. Nwobodo-Nzeribe, "Enhancing trust and transparency in AI: A comprehensive study on explainable artificial intelligence (XAI) techniques and applications," in *Proc. 2024 Int. Conf. Eng. Innov. Sustain. Dev. (ICEISD)*, Enugu, Nigeria, July 2024.

[18] F. Osasona, O. O. Amoo, A. Atadoga, T. O. Abrahams, O. A. Farayola, and B. S. Ayinla, "Reviewing the ethical implications of AI in decision making processes," *Int. J. Manag. Entrep. Res.,* vol. 6, no. 2, pp. 322–335, 2024.

[19] K. Sekiguchi and K. Hori, "Designing ethical artifacts has resulted in creative design: empirical studies on the effect of an ethical design support tool," *AI Soc.,* vol. 36, no. 1, pp. 101–148, 2021.

[20] S. Maleki Varnosfaderani and M. Forouzanfar, "The role of AI in hospitals and clinics: Transforming healthcare in the 21st century," *Bioeng (Basel),* vol. 11, no. 4, pp. 1–38, 2024.

[21] N. A. T. Utami and N. Alawiya, "Perlindungan hukum terhadap pelayanan kesehatan tradisional di Indonesia," *Volksgeist J. Ilmu Huk. dan Konstitusi,* vol. 1, no. 1, pp. 1–17, 2018.

[22] B. Yamini, T. Saraswathi, P. Radhakrishnan, M. Nalini, M. Shanmuganathan, and R. S. Subramanian, "Machine learning algorithms for predicting chronic kidney disease and its significance in healthcare," *Int. J. Adv. Technol. Eng. Explor.,* vol. 11, no. 112, pp. 388–404, 2024.

[23] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning – a brief history, state-of-the-art and challenges," *Commun. Comput. Inf. Sci.,* vol. 1323, pp. 417–431, 2020.

[24] A. Tursunalieva, D. L. J. Alexander, R. Dunne, J. Li, L. Riera, and Y. Zhao, "Making sense of machine learning: a review of interpretation techniques and their applications," *Appl. Sci.,* vol. 14, no. 2, 2024

[25] R. Hoffmann and C. Reich, "A systematic literature review on artificial intelligence and explainable artificial intelligence for visual quality assurance in manufacturing," *Electron,* vol. 12, no. 22, 2023.

[26] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: a survey on methods and metrics," *Electron,* vol. 10, no. 5, pp. 1–19, 2021.

[27] V. Vuppalapaty and T. Architect, "Ethical and legal implications of data sharing in SaaS laboratory management systems," *Asian J. Multidiscip. Res. Rev.,* vol. 5, no. 3, pp. 142–163, 2024.

[28] Y. Okada, Y. Ning, and M. E. H. Ong, "Explainable artificial intelligence in emergency medicine: An overview," *Clin. Exp. Emerg. Med.,* vol. 10, no. 4, pp. 354–362, 2023.

[29] R. Saleem, B. Yuan, F. Kurugollu, A. Anjum, and L. Liu, "Explaining deep neural networks: a survey on the global interpretation methods," *Neurocomputing,* vol. 513, pp. 165–180, 2022.

[30] C. Chen et al., "Artificial intelligence on economic evaluation of energy efficiency and renewable energy technologies," *Sustain. Energy Technol. Assess.,* vol. 47, p. 101358, 2021.

[31] H. A. Bhatt, V. K. Shah, K. Shah, R. Shah, and M. Shah, "State–of–the–art machine learning techniques for melanoma skin cancer detection and classification: a comprehensive review," *Intell. Med.,* vol. 3, no. 3, pp. 180–190, 2022.

[32] A. Martinez-Martinez, J. G. Cegarra-Navarro, A. Garcia-Perez, and A. Wensley, "Knowledge agents as drivers of environmental sustainability and business performance in the hospitality sector," *Tour. Manag.,* vol. 70, pp. 381–389, 2019.

[33] J. Rymarczyk, "Technologies, opportunities and challenges of the industrial revolution 4.0: theoretical considerations," *Entrep. Bus. Econ. Rev.,* vol. 8, no. 1, pp. 185–198, 2020.

[34] V. Narayan and S. Ganapathisamy, "Learning analytics with correlation-based SAN-LSTM mechanism for formative evaluation and improved online learning," *Int. J. Adv. Technol. Eng. Explor.,* vol. 11, no. 112, pp. 373–387, 2024.

[35] B. Vlačić, L. Corbo, S. Costa e Silva, and M. Dabić, "The evolving role of artificial intelligence in marketing: a review and research agenda," *J. Bus. Res.,* vol. 128, pp. 187–203, 2021.

[36] N. Alangari, M. E. B. Menai, H. Mathkour, and I. Almosallam, "Exploring evaluation methods for interpretable machine learning: a survey," *Inf.,* vol. 14, no. 8, pp. 1–29, 2023.

[37] C. Cimini, F. Adrodegari, T. Paschou, A. Rondini, and G. Pezzotta, "Digital servitization and competence development: a case-study research," *CIRP J. Manuf. Sci. Technol.,* vol. 32, pp. 447–460, 2021.

[38] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: applications, challenges and trends," *Neurocomputing,* vol. 408, pp. 189–215, 2020.

[39] Y. Ma and D. Guo, "Exploring the application of big data technology in international trade: case studies from India and China," *J. Logist. Inf. Serv. Sci.,* vol. 10, no. 2, pp. 138–152, 2023.

[40] T. Niebel, "ICT and economic growth – comparing developing, emerging and developed countries," *World Dev.,* vol. 104, pp. 197–211, 2018.

[41] J. Hutson et al., "Artificial intelligence and the disruption of higher education: strategies for integrations across disciplines," *Creat. Educ.,* vol. 13, no. 12, pp. 3953–3980, 2022.

[42] J. C. Bertot and H. Choi, "Big data and e-government: Issues, policies, and recommendations," in *Proc. 14th Annu. Int. Conf. Digit. Gov. Res. (dg.o '13)*, Quebec City, QC, Canada, June 2013, pp. 1–10.

[43] L. J. Moleong, *Qualitative Research Methodology*. Jakarta, Indonesia: PT Remaja Rosdakarya, 2016.

[44] G. Hayashi, P. Abib, and N. Hoppen, "Validity in qualitative research: a processual approach," *J. Qual. Rep.,* vol. 24, no. 1, pp. 98–112, 2019.

[45] H. Bedle and D. L. Robles, "Application of vector plots, LIME, and SHAP for seismic facies machine learning evaluation," in *Proc. 4th Int. Meeting for Applied Geoscience & Energy*, July 2024, pp. 1159–1163.

[46] O. Afolabi, "Balancing Performance and Interpretability in AI Models for Finance and Security," September 2024.

[47] G. Peng, "Segmentation-free recognition algorithm based on deep learning for handwritten text image," *J. Artif. Intell. Technol.,* vol. 4, no. 2, pp. 169–178, 2024.

[48] A. S. Obaid, M. Y. Kamil, and B. H. Hamza, "People recognition via tongue print using deep and machine learning," *J. Artif. Intell. Technol.,* vol. 3, no. 3, pp. 119–125, 2023.

[49] S. Chaudhury and K. Sau, "Classification of breast masses using ultrasound images by approaching GAN, transfer learning, and deep learning techniques," *J. Artif. Intell. Technol.,* vol. 3, no. 4, pp. 142–153, 2023.

[50] M. A. Alafnan, S. Dishari, M. Jovic, and K. Lomidze, "ChatGPT as an educational tool: opportunities, challenges, and recommendations for communication, business writing, and composition courses," *J. Artif. Intell. Technol.,* vol. 3, no. 2, pp. 60–68, 2023.

[51] S. Ali et al., "Explainable artificial intelligence (XAI): what this study know and what has been left to attain trustworthy artificial intelligence," *Inf. Fusion,* vol. 99, p. 101805, 2023.

[52] N. Aslam et al., "Interpretable machine learning models for malicious domains detection using explainable artificial intelligence (XAI)," *Sustain.,* vol. 14, no. 12, p. 7375, 2022.

[53] P. Papadimitroulas et al., "Artificial intelligence: deep learning in oncological radiomics and challenges of interpretability and data harmonization," *Phys. Med.,* vol. 83, pp. 108–121, 2021.

[54] S. Dey and T. R. Chowdhury, "A comparative survey of SHAP and LIME: explaining machine learning models for transparent AI," *Int. J. Innov. Res. Educ.,* vol. 11, no. 6, pp. 827–835, 2024.

[55] A. Gramegna and P. Giudici, "SHAP and LIME: an evaluation of discriminative power in credit risk," *Front. Artif. Intell.,* vol. 4, pp. 1–6, 2021.

[56] A. Cuzzocrea, Q. E. Alahy Ratul, I. Belmerabet, and E. Serra, "Attribution methods assessment for interpretable machine learning," *CEUR Workshop Proc.,* vol. 3478, pp. 65–75, 2023.

[57] S. Knapič, A. Malhi, R. Saluja, and K. Främling, "Explainable artificial intelligence for human decision support system in the medical domain," *Mach. Learn. Knowl. Extr.,* vol. 3, no. 3, pp. 740–770, 2021.

[58] M. Saarela and V. Podgorelec, "Recent applications of explainable AI (XAI): a systematic literature review," *Appl. Sci.,* vol. 14, no. 19, p. 8884, 2024.

[59] M. Ennab and H. Mcheick, "Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions," *Front. Robot. AI,* vol. 11, pp. 1–16, 2024.

[60] A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion,* vol. 58, pp. 82–115, 2020.

[61] E. A. Inigo, L. Albareda, and P. Ritala, "Business model innovation for sustainability: exploring evolutionary and radical approaches through dynamic capabilities," *Ind. Innov.,* vol. 24, no. 5, pp. 515–542, 2017.

[62] G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F. Amenta, "Applications of machine learning predictive models in the chronic disease diagnosis," *J. Pers. Med.,* vol. 10, no. 2, 2020.