Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Explainable Artificial Intelligence for Trustworthy Decision-Making in Critical Domains

Miss. Sanika Narayan Dhande, Student, SYMSc(CA)
Department of Computer Science
Ashoka Center for Business and Computer Studies, Chandsi
Nashik, India

Mr. Rahul Abhiman Sonawane, Assistant Professor
Department of Computer Science
Ashoka Center for Business and Computer Studies, Chandsi
Nashik, India

Abstract - ***The adoption of Artificial Intelligence (AI) and Machine Learning (ML) systems in such vital areas of life as healthcare, finance, and human resource management has enhanced the efficiency and accuracy of decision-making significantly. Most high-performing AI models are however black boxes which provide minimal transparency on the manner in which predictions are made*** [1]***. This uninterpretability is highly questionable in terms of trusting, accountability, fairness and ethical adherence, particularly when the decisions carry the direct effects on the humankind. The lack of explainability can tend to limit the embrace of AI systems by stakeholders and regulators.***

In this study, the problem of model opaqueness is discussed based on the research of Explainable Artificial Intelligence (XAI) methods to improve transparency and trust in AI systems used to make decisions. The experiment uses and tests the most popular XAI techniques including SHAP (Shapley Additive explanations) [2] ***and LIME (Local Interpretable Model-agnostic Explanations)*** [3] ***on machine learning models used on a critical dataset of decisions. These methods are employed to understand which features are important, clarify each individual prediction and give human comprehensible insights into the behaviour of a model.***

The experimental evidence confirms that the technique of integrating the explainability techniques does not have a significant negative impact on the predictive performance, but still has a significant positive impact on the model interpretability. The suggested strategy empowers stakeholders to make informed, verified, and reliable decisions made by AI so as to embrace AI responsibility. The results of this project demonstrate the relevance of explainable models to the implementation of AI systems in sensitive and high-risk settings and future plans on how to combine accuracy, fairness, and transparency in intelligent systems.

Keywords - Explainable AI,Machine learning, Model Interpretability, Trustworthy AI, Critical Decision Systems

I. INTRODUCTION

a) Background

Artificial Intelligence (AI) systems are also being implemented in high stakes settings like medical diagnosis, credit scoring, criminal justice and recruitment. These systems are guaranteed to be efficient, more precise, and scalable. Nevertheless, their extensive use has also been accompanied by the increasing concerns as to their transparency. The predictions of complex models such as deep neural networks and ensemble methods can be regarded as black boxes that are hard to comprehend regarding the justification of the decisions they take *[4]*. Such non-transparency can result in lack of trust in end-users, regulatory issues, ethical dilemmas, especially in areas where the decision-made is related to the well-being and rights of humans. The concept of Explainable Artificial Intelligence (XAI) has come to be the focus of important research in order to make AI systems more readable and responsible *[5]*. XAI techniques can close the performance-to-understandable gap in model decision-making processes to promote a feeling of trust and achieve compliance with legal and ethical conventions like the GDPR regarding the right to explanation *[6]*.

b) Problem Statement

Although AI performance has improved, model interpretability is a major challenge towards their application in sensitive areas. Clinicians, loan officers and regulatory bodies are the main stakeholders who are reluctant to trust the use of AI systems whose decision-making process cannot be elucidated. This issue becomes even worse in the context where the biases, mistakes, or unjust results may be critical. The existing XAI techniques are different in terms of applicability, fidelity, and usability and therefore demand a systematic analysis of their effectiveness in actual practice.

c) Research Objectives

This study aims to:

i. Research how XAI can help to increase the transparency and trust of the AI-driven decision systems.
ii. SHAP and LIME Implement and test on healthcare and financial benchmark datasets.
iii. Determine the trade off between model interpretability and accuracy.
iv. Offer practical recommendations on the implementation of the explainable AI in accordance with ethical and regulatory standards.

d) Scope and Limitations

This paper is based on post-hoc explainability methods of supervised classification models. The data sets are obtained in healthcare and finance as publicly available. Limitations in the research are the access to labeled data and generalizability of results in other domains. Future research can be expanded to real time explainability and unsupervised learning conditions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

II. LITERATURE REVIEW

A. Theoretical Foundations

The explainability requirement of AI is based on the ethical principles of AI, human-centered design, and regulations. Theories like Trust in Automation and Algorithmic Accountability state that users should know and justify automated decisions to trust them. In the same vein, the GDPR requiring model interpretability has rendered model interpretability legally binding in numerous jurisdictions such as the Right to Explanation.

B. Previous Research

The different XAI methods have been examined in the past:

- LIME (Ribeiro et al., 2016) is a technique that comes up with local explanations through approximating complex models using interpretable ones [3].
- Games theory is applied in SHAP (Lundberg and Lee, 2017) to assign the significance of features in a regular way [2].
- In healthcare (e.g., Caruana et al., 2015), it has been demonstrated that the explainable models can enable clinicians to trust AI-assisted diagnoses [7].
- Research also demonstrates that explainable credit scoring models enhance customer satisfaction and regulation in the financial industry [8].
- Nevertheless, there are spaces of evaluation of the practical utility of these approaches in various critical areas and the perception of what effects they have on user confidence and decision-making [9].

C. Gaps in Current Research

- Absence of comparative literature on local and global explainability methods [10].
- Lack of emphasis on the implementation of XAI in the current decision-making processes.
- Very little studies have investigated the longitudinal effect of explainability on user trust and adoption [11].

III. METHODOLOGY

A. Research Design

The research is based on the experimental design, which concurrently incorporates measurement of quantitative performance of the model and qualitative measure of the explainability. The study is carried out in three stages:

1) Training and validation of model training.
2) SHAP and LIME techniques of XAI.
3) Comparison on basis of accuracy measures and interpretability measures.

B. Datasets

There are two publicly available datasets that are utilised:

- Healthcare Heart Disease UCI Dataset.
- German Credit Data (Finance).

They are both binary classification using features pertaining to critical decision-making.

C. Models and Tools

- Deep Neural Network, Gradient Boosting, Random Forest.
- XAI Tools SHAP (python library), LIME (python library)
- Measures of Evaluation: Accuracy, Precision, Recall, F1-Score, AUC-ROC; Explanation Fidelity, Consistency and User Comprehension Scores.

D. Implementation Steps

1) Pretraining and training models.
2) Make explanations based on SHAP (global and local) and LIME (local).
3) Comparison of model action with and without explanation.
4) The survey of stakeholders (N=20) to determine the clarity and usefulness of the explanations is conducted.

IV. SYSTEM DESIGN / CONCEPTUAL FRAMEWORK

A. Framework Overview

It suggests a conceptual framework of integrating XAI as a critical decision system, which includes:

1. Data Layer- Processed, domain specific data.
2. Model Layer - ML models of high performance.
3. Explainability Layer Rules that explain the model.
4. interface Layer- Stakeholder-friendly dashboards.
5. Validation Layer- Networking with moral and regulatory regulations.

B. Component Integration

The framework makes sure that real-time explanations are created, displayed on user-friendly formats (visualizations, textual summaries), and registered to be audited. This leads to transparency, eases debugging and builds user trust.

V. EXPERIMENTAL RESULTS AND ANALYSIS.

A. Model Performance

Model	Accuracy (%)	F1-Score	AUC-ROC
Random Forest	91.2	0.89	0.94
Gradient Boosting	89.7	0.87	0.92
Deep Neural Network	88.5	0.86	0.91

Table 1: Predictive results on Heart Disease data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

B. Explainability Evaluation

XAI Method	Fidelity Score	Consistency	User Comprehension (Avg. Score /10)
SHAP	0.92	High	8.5
LIME	0.88	Medium	7.8

Table 2: Measures of explainability.

C. Stakeholder Feedback

Findings of the surveys showed that:

- Shareholders rated SHAP summaries 85 percent useful in grasping the general model behavior.
- Three-quarters (78%) found MIKE more favorable in case-specific explanations.
- 90% of them noted that they had more trust in AI decisions when they were explained.

VI. PRIVACY, SECURITY, AND ETHICAL CONSIDERATIONS

A. Data Privacy

All data sets were anonymized and utilized according to the regulations of data protection. The explanations were made in such a way that there would not be leakage of sensitive information.

B. Ethical Implications

The paper highlights the significance of:

- Not biased explanations [12].
- Promoting equity in model interpretations [13].
- Giving straightforward explanations to the non-expert users.

C. Regulatory Alignment

The suggested XAI integration is in line with GDPR, HIPAA, and future AI ethics standards and advocates accountability and transparency [6,16].

VII. DISCUSSION AND CONCLUSION

A. Interpretation of Findings

The findings affirm that XAI methods can be highly effective to improve the interpretability of models without affecting the precision of the results [14]. SHAP was more consistent in global insights and LIME was more local, instance-level. Explainability provided more trust between stakeholders and made it easier to validate models.

B. Comparison with Existing Work

This research builds upon the work of other researchers by:

- Providing a comparative study of SHAP and LIME on important areas [15].
- Suggesting a deployable explainable AI system framework.
- Adding user-feedback to determine usefulness.

C. Conclusion

Explainable AI is not a technical improvement but a necessity of reliable AI in areas of concern [1,5]. This study proves that approaches such as SHAP and LIME have the potential to reduce the distance between complex models and human cognition and make AI responsible.

Future work should focus on:

- Explainability in a dynamic environment in real-time.
- Visual explanations (explanations through text, descriptions, visual, interactive).
- Explainability assessment measures which are standardized [17].

VIII. ACKNOWLEDGMENTS

The author appreciates the open-source communities of SHAP and LIME, their tools being of great help. No external funding was accorded to this research.

REFERENCES

[1] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138-52160.

[2] Arrieta, A. B.-R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

[3] Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning.* Cambridge, MA: MIT Press.

[4] Caruana, R. L. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). New York, NY: ACM.

[5] Doshi-Velez, F. &. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*.

[6] European Commission. (2019). *Ethics Guidelines for Trustworthy AI.* Brussels, Belgium: European Commission.

[7] Goodman, B. &. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine, 38*(3), 50–57.

[8] Gunning, D. &. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine, 40*(2), 44–58.

[9] Hall, P., Gill, N., & Kurka, M. (2019). *Machine learning interpretability with H2O driverless AI.* Mountain View, CA: H2O.ai.

[10] Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. *AAAI/ACM Conference on AI, Ethics, and Society* (pp. 131-138). Honolulu, HI: Association for Computing Machinery.

[11] Lundberg, S. M. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30 (NeurIPS)* (pp. 4765–4774). San Diego, CA: Neural Information Processing Systems Foundation.

[12] Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). Munich, Germany: Self-published / Online.

[13] Ribeiro, M. T. (2016). Why should I trust you?" Explaining the predictions of any classifier. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). San Francisco, CA: Association for Computing Machinery (ACM).

[14] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206-215.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

[15] Samek, W. W. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv*.

[16] Stepin, I., Alonso, J. M., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access, 9*, 11974-12001.

[17] Zhang, J., Liao, Y., & Wang, S. (2023). A survey on evaluation of explainable AI. *ACM Computing Surveys, 55*(9), 1-38.