Original Article

Check for updates

Available online at
https://jiicet.gnt.com.pk/index.php/jiicet/

Journal of Innovative Intelligent Computing and Emerging Technologies (JIICET)

Vol. 1, No. 1, 2026

Bridging the Trust Gap: Integrating Explainable AI with IoT-Enabled Machine Learning Systems

Sayed Mazhar Ali[1], Mushtaque Ahmed Rahu[2]

[1] Department Mechatronics Engineering, Air University Karachi, Pakistan, email: mazharlakyari@gmail.com (S.M.A)
[2] Department of Electronic Engineering, QUEST Nawab Shah, 67450, Pakistan, email: rahumushtaque@gmail.com (M.A.R)

* Correspondence: mazharlakyari@gmail.com

Article History

Received 30 December 2025
Revised 29 Jan 2026
Accepted 02 April 2026
Available Online 08 April 2026

Keywords:
Explainable AI (XAI),
Internet of Things (IoT)
Machine Learning (ML)
Trustworthy AI
Edge Computing, Model

Abstract

Due to the combination of the Internet of Things (IoT) and Machine Learning (ML), it has developed effective predictive and automated networks. Nonetheless, transparency, trust, and regulatory compliance of advanced ML models, especially deep learning, is highly challenging due to its black-box nature of operation, especially in high-stakes IoT applications. The paper discusses how it is possible to incorporate Explainable AI (XAI) procedures to improve the explanation and responsibility of ML models used in IoT systems. We compare the most recent XAI methods applicable to IoT-ML pipelines, examine how they can be applied in smart healthcare and industrial IoT, and discuss a framework to be employed. Such issues as the computational overhead, the generation of explanations in real-time, and the measurement of the evaluation are also open challenges discussed in the paper. Through a literature review of the recent literature, we present the argumentation that XAI is not an add-on to the deployment of intelligent IoT systems but an essential element to their sustainable and ethical deployment.

1. Introduction

The Internet of Things (IoT) creates enormous, heterogeneous data streams of the billions of sensors and devices that are interconnected. Machine Learning particularly deep neural networks is good at deriving patterns and insights out of this data allowing applications in predictive maintenance and smart grids to customized healthcare [1]. The complexity and obscurity of high-performance ML models are, however, operational and therefore form a trust deficit. The stakeholders, including engineers, end-users and regulators, may find it difficult to know how or why a model made a particular decision, including raising a flag on a piece of industrial equipment to indicate failure or testing a medical condition [2].

Explainable AI (XAI) is a key area that has come up to counter this opaqueness. It includes methods of rendering the behaviour and outputs of AI systems comprehensible to humans [3]. In the case of IoT-ML systems, which are commonly used in safety-critical and privacy-sensitive settings, XAI is not a desirable attribute anymore but a requirement. In this paper, the intersecting area of these three areas is explored. IoT provides the data substrate, ML delivers predictive capabilities, and XAI ensures transparency and trust in the loop as shown in Figure 1.

The main issue is the tension between the complexity-related model performance (most of the time) and interpretability in IoT-deployed ML. This research aims to:

1. Survey state-of-the-art XAI methods in relation to IoT-ML systems.

2. Discuss the difficulties in implementing XAI under the limits of IoT (e.g. edge computing, bandwidth, latency).

3. Suggest a light framework of incorporating XAI into the pipelines of the Internet of Things-ML.

4. Determine the future research directions by gaps in existing literature.

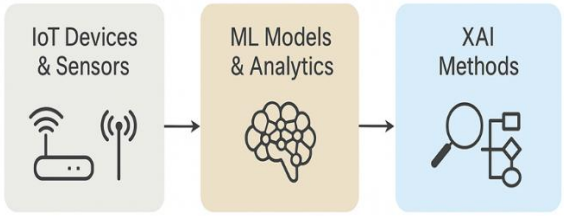

Figure 1. The convergence of XAI, IoT, and ML

2. Background and Related Work

i. IoT and ML Synergy

The IoT systems are defined with the 5Vs that are Volume, Velocity, Variety, Veracity, and Value. ML algorithms are used to detect anomalies, make predictions in a time series as well as classification on this data stream [4]. A common type of architecture is to use sensor ingested data, pre-process, infer a model of the sensor data (at the edge or in the cloud), and actuate.

ii. The Need for XAI in IoT

The demand for XAI in IoT stems from multiple factors:

Safety & Accountability: In autonomous vehicles or medical IoT, understanding failure modes is crucial.

Regulatory Compliance: Regulations like the EU AI Act mandate transparency for high-risk AI systems [5].

Model Debugging & Improvement: Explanations help developers identify biases, data drift, or flawed logic.

User Trust & Adoption: Clear explanations foster acceptance of AI recommendations by operators or clinicians.

Surveys undertaken in recent times have started mapping this terrain. A review of XAI to the security of IoT was made in [6], whereas deep learning to interpretability techniques were studied in [7]. Nevertheless, an in-depth study of the system-level integration, taking into account the distributed character of the IoT, has not yet been studied thoroughly.

3. XAI Techniques for IoT-ML Systems

The XAI techniques can be classified as intrinsic (that is, the ones understandable by design) or post-hoc (that is, the ones that are applied after the model is trained). Post-hoc methods are commonly used because of the complexity of the models that may be required in the case of the IoT data.

i. Model-Agnostic Post-Hoc Methods

These are flexible and can be applied to any ML model.

SHAP (SHapley Additive exPlanations): According to the cooperative game theory, SHAP uses the value of importance of each feature in a given prediction [8]. It is efficient and potentially costly to compute.

LIME (Local Interpretable Model-agnostic Explanations): Estimates complex model in the locality with an interpretable one (e.g. linear model) to describe predictions (individual) [9].

Counterfactual Explanations: Explain the minimal alteration that should be made to the input features in order to modify the model output (e.g. your loan was rejected due to your income being €5K below the level). By raising it to €XK, they would be allowed to approve it.) [10].

ii. Model-Specific Methods

Attention Mechanisms: In the case of sequence models (e.g. sensor data analysis), attention layers can be used to visualize what the model is "attentive to" in an input sequence to make a decision [11].

Grad-CAM (Gradient-weighted Class Activation Mapping): It is mostly used by CNNs to highlight significant areas in an image (or other points of interest in structured data) upon which to make a prediction [12].

iii. Selecting XAI for IoT Constraints

Such strategies have to be implemented in IoT systems with the consideration of computational resources, latency and energy. Table 1 makes a comparison of major techniques.

Table 1. Comparison of XAI Techniques for IoT-ML Deployment

XAI Technique	Interpretability Scope	Computational Cost	Suitability for IoT Edge	Primary Output
SHAP (Kernel)	Global & Local	Very High	Low (Cloud)	Feature importance values

SHAP (Tree)	Global & Local	Medium	Medium (Powerful Edge)	Feature importance values
LIME	Local	Medium	Medium (Powerful Edge)	Local linear model coefficients
Counterfactuals	Local	Medium-High	Low-Medium	"What-if" input examples
Attention Weights	Local	Low (built-in)	High (Lightweight)	Attention heatmaps over sequences
Rule-based Models	Global	Low	High (Very Lightweight)	Human-readable IF-THEN rules

4. An Integrated Framework for XAI in IoT-ML

Our suggestion is a hierarchical model of XAI integration into an IoT-ML pipeline, as shown in Figure 2. Architecture depicts the data acquisition to insights that can be acted upon and explained, where XAI modules can be found at the edge and cloud layers.

Layer 1: Data & Sensing Layer. Raw data is collected from heterogeneous IoT sensors.

Layer 2: Edge Processing Layer. Portable data processing and filtering take place. This can be used to run lightweight XAI modules (e.g., attention visualization of RNNs, very simple rule extractors) in real-time, low-latency to make explanations about significant events. There can also be model inference of time sensitive tasks.

Layer 3: Fog/Cloud Layer. Other ML models that are more massive are trained and deployed. Here contains a centralized XAI Engine which does more detailed and computationally expensive explanation generation (e.g., global SHAP, detailed counterfactuals) when requested or during a batch analysis.

Layer 4: Explanation Interface Layer. Both edge and cloud explanations will be presented in a format that suits various stakeholders (e.g., the explanations will be presented in JSON format to a developer dashboard, as a natural language summary to operators, and as visual highlight to clinicians).

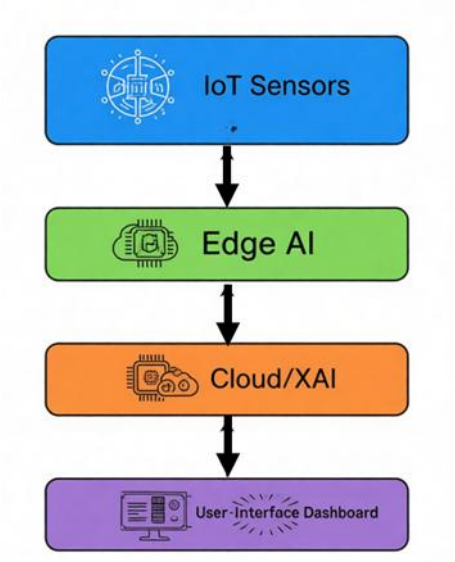

Figure 2: Proposed XAI-IoT-ML Integration Framework

i. Case Study: Predictive Maintenance in Industrial IoT

Scenario: The data is provided to a deep learning model that predicts risk of failure based on vibration sensors installed on the motors.

ML Model: A time-series classification convolutional neural network (CNN) in 1D.

XAI Integration: On the edge, a simplified version of Grad-CAM is used to determine which time-window of the sensor signal best predicted failure. SHAP analysis in the cloud is performed every week to prioritize the value of various sensor types and conditions of operation on the whole fleet.

Outcome: The maintenance engineer is informed: in 48 hours, Motor-7 has a 92 percent probability of failure. It is mainly because of abnormal high-frequency vibration pattern (highlighted by red), which was detected between 14:30-14:45, and is consistent with historical bearing failure signatures.

5. Challenges and Future Research Directions

i. Persistent Challenges

Computational Overhead vs. Utility: Most high-fidelity XAI approaches are prohibitively expensive in terms of resource consumption by edge devices. Studies of ultra-lightweight, approximate methods of explanation are required.

Explanation Consistency & Evaluation: The amount of gold to measure the goodness of an explanation does not exist. Such measures as faithfulness (is the explanation the actual reasoning of the model) and stability (do similar inputs give similar explanations?) remain subject to ongoing research [13].

Real-Time Explanation Generation: Producing useful explanations within milliseconds of getting a high-velocity IoT stream is still a challenge.

Human-Centric Design: The user will need explanations that are based on his or her expertise (data scientist or field technician). The visualization and presentation of explanation in the context of the IoT is poorly developed [14].

ii. Promising Future Directions

Inherently Interpretable Models for IoT: The further evolution of high-performing and intrinsically interpretable models (e.g., new types of rule-based systems, concept bottleneck models) adapted to the data of the IoT [15].

XAI-aware Edge AI Hardware: Creating next-generation AI accelerators (e.g., TPUs, NPUs) with implicit hardware support of effective XAI computing.

Federated XAI: Generalizing federated learning to produce explanations in privacy-assuring, decentralized IoT networks without centralizing raw data [16].

Causal XAI: That is why the correlational explanations can be replaced by causal explanations that determine cause-effect relationships in IoT systems to make it easier to make decisions [17].

6. Conclusion

The new era of digital transformation is the integration of IoT and ML, and the concern about transparency and trust is reasonable. Explainable AI can help to supply the necessary toolkit to handle these issues. In this paper, it was argued that, XAI should be a fundamental architectural feature not an add-on in IoT-ML systems. We revved relevant XAI methods, suggested an integrative framework which considers constraints of the IoT, and identified key challenges.

The deployment of the successful deployment takes the cooperation of the ML researchers, integrated systems engineers and HCI specialists. Due to the incessant changes in standards and regulations, the capability to create actionable and comprehensible explanations will also become a focal point of responsible and successful AI-based IoT solutions. Future research ought to involve the standardization of assessment benchmarks on XAI in IoT and creation of resource efficient explanation techniques to be able to run at the extreme edge.

Author Contributions: "Conceptualization, M.A.R. and S.M.A.; methodology, M.A.R; and S.M.A writing—original draft preparation, M.A.R.; writing—review and editing, S.M.A.; visualization, M.A.R.; supervision, S.M.A.; All authors have read and agreed to the published version of the manuscript."

Funding: "This study does not receive external funding."

Ethical Clearance: "Not applicable".

Informed Consent Statement: "Not applicable".

Data Availability Statement: "Not applicable".

Acknowledgments: Thanks Co-author for support and help.

Conflicts of Interest: "All the authors declare that there are no conflicts of interest."

References

1. M. S. Mahdavinejad et al., "Machine learning for internet of things data analysis: A survey," Digital Communications and Networks, vol. 4, no. 3, pp. 161-175, 2018.
2. A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, 2020.
3. D. Gunning et al., "XAI—Explainable artificial intelligence," Science Robotics, vol. 4, no. 37, eaay7120, 2019.
4. J. Qiu et al., "A survey of machine learning for big data processing," EURASIP Journal on Advances in Signal Processing, vol. 2016, no. 1, p. 67, 2016.
5. European Commission, "Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," 2021.
6. M. A. Ferrag et al., "Explainable deep learning for cybersecurity in IoT networks," IEEE Access, vol. 10, pp. 93104-93139, 2022.
7. A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," arXiv preprint arXiv:2006.11371, 2020.
8. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in neural information processing systems, vol. 30, 2017.
9. M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144, 2016.
10. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harv. JL & Tech., vol. 31, p. 841, 2017.

11. A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
12. R. R. Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," Proceedings of the IEEE international conference on computer vision, pp. 618-626, 2017.
13. Y. R. Asaithambi and K. K. R, "A survey on evaluation of explainable artificial intelligence methods," International Journal of Information Technology, pp. 1-9, 2023.
14. H. Lakkaraju et al., "Faithful and customizable explanations of black box models," *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 131-138, 2019.
15. C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: deep learning for interpretable image recognition," Advances in neural information processing systems, vol. 32, 2019.
16. L. Lyu et al., "Privacy and robustness in federated learning: A survey," IEEE Transactions on Knowledge and Data Engineering, 2022.
17. J. Pearl, "The seven tools of causal inference, with reflections on machine learning," Communications of the ACM, vol. 62, no. 3, pp. 54-60, 2019.
18. M. Dangrach, S. M. A. Shah, A. Z. ul Abdin, A. Ali, and M. Ahmed, "IoT based fire system," Journal of Applied Engineering & Technology, vol. 5, no. 1, pp. 19–30, 2021.
19. L. B. Gokalani, B. Das, D. K. Ramnani, M. Kumar, and M. A. Shah, "House price prediction of real-time data (DHA Defence Karachi) using machine learning," Sir Syed University Research Journal of Engineering & Technology, vol. 12, no. 2, pp. 75–80, 2022.
20. M. A. Rahu, A. F. Chandio, and S. M. Ali, "Machine learning overview in agriculture," Journal of Applied Engineering & Technology, vol. 6, no. 1, pp. 28–39, 2022.
21. G. M. Jatoi, M. A. Rahu, S. Karim, S. M. Ali, and N. Sohu, "Water quality monitoring in agriculture: Applications, challenges and future prospects with IoT and machine learning," Journal of Applied Engineering & Technology, vol. 7, no. 2, pp. 46–54, 2023.
22. R. M. Ahmed, "Integration of wireless sensor networks, Internet of Things, artificial intelligence, and deep learning in smart agriculture: A comprehensive survey," Journal of Innovative Intelligent Computing and Emerging Technologies, vol. 1, no. 1, pp. 1–9, 2024.
23. M. A. Sayed, "The internet of things (IoT), applications and challenges: A comprehensive review," Journal of Innovative Intelligent Computing and Emerging Technologies, vol. 1, no. 1, pp. 20–27, 2024.
24. M. A. Rahu, M. M. Shaikh, S. Karim, S. A. Soomro, D. Hussain, and S. M. Ali, "Water quality monitoring and assessment for efficient water resource management through IoT and machine learning approaches for agricultural irrigation," Water Resources Management, vol. 38, no. 13, pp. 4987–5028, 2024.
25. S. M. A. Shah, A. Soomro, K. Hussain, M. A. Rahu, and S. Karim, "Innovative machine learning solutions: Investigating algorithms and applications," Journal of Applied Engineering & Technology, vol. 8, no. 1, pp. 32–51, 2024.
26. S. M. Ali, B. Das, and D. Kumar, "Machine learning based crop recommendation system for local farmers of Pakistan," Revista Geintec – Gestão, Inovação e Tecnologias, vol. 11, no. 4, pp. 5735–5746, 2021.
27. M. A. Rahu, M. M. Shaikh, S. Karim, A. F. Chandio, S. A. Dahri, S. A. Soomro, and S. M. Ali, "IoT and machine learning solutions for monitoring agricultural water quality: A robust framework," Mehran University Research Journal of Engineering & Technology, vol. 43, no. 1, pp. 192–205, 2024.
28. B. Das, S. M. Ali, M. Z. Shaikh, A. F. Chandio, M. A. Rahu, J. K. Pabani, and M. U. R. Khalil, "Linear regression based crop suggestive system for local Pakistani farmers," in Proc. 2023 Global Conf. on Wireless and Optical Technologies (GCWOT), IEEE, 2023.
29. S. M. Ali, M. A. Rahu, S. Karim, and S. A. Soomro, "Ensuring security and privacy in Internet of Things deployments for industry, training, and residential environments: A comprehensive investigation," in Transforming Industries: Capturing the Potential of IIoT and ML in the Era of Industry 5.0, CRC Press, pp. 115–126, 2025.