International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(22s) 2026 pp. 515-525
https://doi.org/10.70917/ijcisim-2026-5457
Article

Explainable AI for IoT-Driven Decision Support: Building Trust in Human-AI Collaboration via Deep Learning Interpretability

Faisal Rahman[1], Ayesha Anzer[2], Ameer Hamza Nawaz[3], Zahoor Ahmed[4], Hafiz Tasawar Hussain[5], Shoaib Hayat[6], Zaid Wali[7], Nazia Azim[8]

[1] Departmental Information Technology Management, Southwest Baptist University MO USA.
Email: faisal.rahman@sbuniv.edu
[2] Faculty of Engineering and Applied Science, University of Regina, SK S4S 0A2, Canada.
Email: aag833@uregina.ca
[3] MS Scholar, HITEC University Taxila, Pakistan.
Email: nawazhamza464@gmail.com
[4] Department of Electrical Engineering, Balochistan University of Engineering and Technology, Khuzdar, Pakistan.
Email: zahoor@buetk.edu.pk
[5] Lecturer, Faculty of Computer Science and Information Technology, The Superior University, Lahore, Pakistan.
Email: mateentasawar80@gmail.com
[6] Department of Computer Science and Technology, Auckland University of Technology, (AUT) Auckland, New Zealand.
Email: shoaib.devpy@gmail.com
[7] Abasyn University Islamabad.
Email: zaidwali73@gmail.com
[8] Lecturer Department of Computer Science Abdul Wali Khan, University Mardan, Pakistan.
Email: N.azim@awkum.edu.pk
Corresponding Author: Zaid Wali

Abstract: This study was focused on the development and assessment of an explainable artificial intelligence (XAI) system for an Internet of Things (IoT) based decision support system, not only in terms of prediction accuracy, but also in how it impacts human-AI collaboration. A computational experimental methodology was used in which, IoT sensor data were gathered, cleaned, normalized, treated for missing values, and selected features, and then partitioned into training, validation, and testing datasets. A deep learning model was developed to make predictions that supported decision making; subsequently, model agnostic explanation techniques were embedded, post hoc, to uncover and feed-back the most salient features that lead to each prediction. Predictive performance was assessed using accuracy, precision, recall, F1-score and area under the receiver operating characteristic curve compared to an existing deep learning model without explanatory mechanisms. The consistency and the relevance of generated explanations were then evaluated and 160 participants were then exposed to AI-driven decisions with or without explanations to test the impact on trust, perceived understanding, decision confidence, and willingness to trust the system. The findings showed that the explainable model can predict as well as the conventional model, the explanations produced by the two different interpretability techniques were very similar, and participants who received the explanations of the decisions they made were significantly more confident, understood, trusted and willing to rely on the decision made by the explainable model than those who were not given the explanations. Further, the quality of explanations was also determined to be significantly and positively correlated with user trust. The results indicate that interpretability can be integrated into the deep learning-based decision support for IoT without compromising its predictive power and effectively enhance the integration between humans and AI.
Keywords*:* explainable artificial intelligence, Internet of Things, deep learning, interpretability, human-AI collaboration, user trust

1. Introduction

With the prevalence of Internet of Things (IoT) sensors in industrial, healthcare and infrastructure applications, there have been an abundance of real-time operational data that greatly inform the automation of decision making (Al-Fuqaha et al., 2015). To leverage this large and complex data has led organizations to deep learning

models which have shown to outperform traditional methods of statistics and machine learning in many IoT-based applications (LeCun et al., 2015).

Although powerful, deep learning models have been criticized as black boxes and have been known to produce correct predictions without explanation for the prediction. (Rudin, 2019) For an IoT-based decision support scenario, predictions can trigger significant operational and safety decisions, and humans might need to be able to account for and approve the decisions generated by the system before taking any measures to implement them (Samek et al., 2017).

To address this challenge, a new concept has emerged, called explainable artificial intelligence (XAI), which involves a variety of methods aimed at making the inner workings of complex models more comprehensible to human users (Adadi & Berrada, 2018). The model-agnostic methods like SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) have enabled the quantification of individual input feature contribution, and visualization-based methods, such as Grad-CAM (Selvaraju et al., 2017), have made deep convolutional networks interpretable.

Incorporating explainability into IoT-based decision support goes beyond a mere technical task; it is a matter of how AI systems and humans will work together. The true power of explainability lies in its ability to foster trust, understanding, and safe reliance on the system, which is a critical aspect of human-AI collaboration (Miller, 2019). Previous studies on human trust in automation have reported that when the technical level of explainability is low, people can be distrustful of or algorithm averse towards such systems, even if they perform as well as or better than humans in making predictions (Dietvorst et al., 2015), highlighting the importance of examining explainability both technically and impactfully on human users.

While a number of studies have proposed and benchmarked XAI techniques based on predictive performance (Doshi-Velez & Kim, 2017; Guidotti et al., 2018), far fewer studies have integrated predictive performance metrics of a predictive XAI model for IoT data and the subsequent impact of the explanation on human trust and collaboration outcomes (Arrieta et al., 2020). This is a significant gap, because the explanatory method must actually enhance the human-AI collaborative relationship that it is designed to support in order to be of practical value (Ashoori & Weisz, 2019).

In this context, the present study introduces and assesses an explainable AI system for decision support using an IoT service, which includes a predictive model based on deep learning, model-agnostic interpretability methods, and a human-centered trust and collaboration outcome assessment. The study meets the desire for more interpretable deep learning architectures that can enable real-time decision-making in the IoT by empirically evaluating whether interpretability, besides predictive accuracy, meaningfully enhances human-AI collaboration (Mohammadi et al., 2018). Specifically, it aimed to develop and test an explainable deep learning model for IoT-based decision support and compare the quality of explanations it produced with that of a traditional, non-explainable deep learning model; and to test whether the quality of explanations was a significant predictor of trust, perceived understanding, decision confidence, and willingness to rely on the system.

2. Literature Review

Decision Support Systems Driven by IoT

Continuous streams of sensor data are essential to enable decision making for operation, diagnostics or safety (Al-Fuqaha et al., 2015) in an IoT-based decision support system. Atzori et al., (2010) defined IoT as a synthesis of sensing, communication and computation technologies that can convert "information from the environment" into "information to act", in other words. Industrial and supply chain applications have focused on the use of IoT for decision support to identify equipment anomalies, predict equipment failures and improve resource allocation, before the issues become costly disruptions (Ben-Daya, Hassini, & Bahroun, 2019). The growing number, speed, and scale of the IoT data has necessitated the increased need for deep learning to be able to infer predictions from sensor streams, making deep learning a natural fit for the computational backbone of IoT-driven decision support, as observed by Mohammadi et al., (2018).

The IoT data is processed using deep learning.

Multi-layered models of representation learning are termed deep learning models, and have shown significant gains in predicting image, text, and sensor data over traditional machine learning models (LeCun, et al., 2015). According to Goodfellow et al., (2016) the performance improvement is due to deep architectures' ability to automatically learn the hierarchical feature representations directly from scratch, thus eliminating the need for manual

feature engineering. In the context of IoT, deep learning has been used for predictive maintenance, anomaly and intrusion detection, and many other applications, which, in many cases, can be better than shallow models for data with high dimensionality and noise in the sensor data domain (Mahdavifar & Ghorbani, 2019). In the healthcare domain, Esteva et al. (2019) showed that deep learning models could match and even surpass the diagnostic accuracy of human experts, and that this trend is likely to continue spreading to other high-stakes, data-rich domains in which experts are making decisions, such as those in the IoT sector.

The Black-Box Problem and the Case for Explainability

Deep learning models often lose interpretability in the process of learning the predictive relationship, through the application of many nonlinear transformations across the hidden layers (Rudin, 2019). Lipton (2018) distinguished between transparency (how easily the mechanics of a model can be understood) and post hoc interpretability (how easily the mechanics of a model can be explained after a prediction has been made), claiming that most deep learning models are only post hoc interpretable. This opacity becomes an issue of particular concern in domains in which flawed predictions have real operational or safety implications, because the user has no way of knowing whether a model's reasoning is correct or just a lucky guess, Samek et al., (2017). Beyond its academic interest, Molnar (2020) also pointed out that the need for interpretability is not only of a practical, ethical, and regulatory nature, but calls for machine decisions to be made interpretable to the humans impacted by them.

Explainable AI Techniques

There are various methods that can be used to explain a complex model, known as explainable AI (XAI) techniques. Ribeiro et al., (2016) proposed an approach called LIME which explains individual predictions using an interpretable surrogate model that approximates the local decision boundary of a complex model. The SHAP approach, proposed by Lundberg and Lee (2017) and based on cooperative game theory, uses a contribution value that quantifies the marginal contribution of each input feature to the prediction, and provides theoretically grounded properties that were not shared by many previous methods. Selvaraju et al., (2017) created Grad-CAM, a visual method that employs gradient information to generate heatmaps of the parts of an input that are most important to a model's prediction, particularly for deep convolutional architectures. In their extensive review of interpretability methods, Guidotti et al., (2018) classified these methods as either model-specific or model-agnostic, and either individual prediction or model behaviour-explaining methods, and finally concluded that none of the methods was always better than the others in all contexts. Doshi-Velez and Kim (2017) also noted that the methodological development of evaluation of interpretability methods is underdeveloped and suggested a taxonomy of application-grounded, human-grounded and functionally-grounded evaluation approaches. In their paper, Montavon, Samek, and Müller (2018) have surveyed approaches specifically developed for deep neural network decision explanations, and highlight the importance of evaluation criteria, one of which is frequently overlooked – the degree to which the explanation is faithful to the actual reasoning of the underlying model, which is known as the fidelity of the explanation.

Human-AI Systems: Human Trust and Collaboration

Human trust in an AI system has been defined as a willingness to accept the output of the system, which can only be developed by repeated interaction with the system and cannot be extended automatically, based on positive expectations of the system's behavior (Hoff & Bashir, 2015). Parasuraman and Manzey (2010) warned that overtrust or mistrust of automation can lead to failure in human-automation teaming; overtrust leading to complacency and lack of trust leading to disuse of otherwise reliable systems. Previous research has provided empirical support for algorithm aversion, as users are prone to mistrust algorithmic outputs after a minor error in the output, even in the face of the algorithm's overall accuracy (Dietvorst et al., 2015). In their review of empirical studies on human trust in AI, Glikson and Woolley (2020) found that transparency and explainability of a system are among the factors most consistently found as antecedents of human trust in AI, with explainability identified as a key lever in enhancing human-AI collaboration.

Trust and Collaboration.

However, an increasing number of studies have directly investigated the relationship between explainability and trust and collaboration outcomes compared to non-explainable systems. To be effective, explanations must be selective, contrastive, and socially contextualized, rather than simply a list of potential contributing factors, Miller (2019) suggested. Recently, Zhang, Liao, and Bellamy (2020) observed that the quality of explanations, along with communicated model confidence, greatly affected both the accuracy of the users' decisions and the degree of user trust in AI-assisted decision making. Shin (2021) also showed that both explainability and causability had a significant influence on both user trust and behavioral intention, and that explanations that were not designed properly failed to

elicit the expected benefits of user trust. The explanation-Trust relationship is not linear, and Kizilcec (2016) noted that over-explanation can lead to diminishment of user trust in the system, as well as to closure of the loop.The relationship between explanation and trust is not linear, and too much explanation can lead to a reduction of user trust in the system, as well as closure of the loop (Kizilcec, 2016). Furthermore, in their article (2019) Ashoori and Weisz included explanation quality among the factors that influenced whether users trusted an AI-enhanced decision-making process or not, and whether it led to trustful collaboration or not, demonstrating that explainability should be assessed directly in relation to trust and collaboration results, not taken for granted.

Theoretical Framework

The current study was guided by the responsible artificial intelligence approach, which suggests that artificial intelligence systems designed for consequential knowledge systems should meet technical performance criteria as well as requirements for interpretability, fairness and accountability to be trusted by the humans that depend on them (Arrieta et al., 2020). In this context, model performance and explainability are not at odds with one another, and the real test of the value of an xAI system is its impact on the trust, understanding, and collaborative decision-making of humans, consistent with models of trust in technology that focus on the direct experiences and cognitions of users interacting with a system (McKnight et al.,2011). The study employed two evaluation approaches: a computational assessment and a trust assessment to the human user, in line with this theoretical position.

Hypotheses Development

From the theoretical and empirical review above, the following hypotheses are made:

H1; The explainable deep learning model exhibits predictive accuracy that is comparable to that of a traditional deep learning model that is not explainable.

H2: The results indicate that when the decision is made by an AI and also explained to the user, there is a significant difference in user trust, understanding, confidence in the decision, and willingness to use the AI compared to when the decision is given without an explanation.

H3: The results indicate that there is a significant and positive correlation between explanation quality and user trust.

3. Methodology

Research Design

A computational experimental approach was taken to design and test an explainable artificial intelligence (XAI) based framework for IoT based decision support. The methodology involved two parts: a technical phase (model development) during which a deep learning model and ancillary interpretability mechanisms were developed and evaluated, and a human phase (trust and collaboration responses to explained and unexplained AI decisions) during which a human experiment was conducted.

Collect and describe IoT

A long term monitoring environment of the operational condition of an industrial site was continuously monitored using IoT sensor data such as temperature, vibration, pressure and equipment-status readings. The resulting dataset included 42,000 time-stamped sensor readings for 18 input features, each accompanied by a label corresponding to the operational outcome that followed, a typical setup for IoT predictive-maintenance and anomaly-detection problems.

Data Preprocessing

The gathered data from the IoT systems were preprocessed using a defined data pipeline before the model development. Duplicate and corrupted sensor readings were excluded from the data via data cleaning, and sensor reading variables that lacked readings (about 3.2% of observations) were imputed using k-nearest-neighbor imputation. The continuous features were scaled on a min-max basis to the interval [0, 1] to make them comparable, given that the various sensors had different measurement units. Then, a feature selection process was carried out based on the correlation-based filtering and recursive feature elimination, which led to a feature set of 14 predictors to be retained for model development from the initial 18. The preprocessed dataset was then divided into training (70%), validation (15%) and testing (15%) sets using stratified random sampling to ensure that the class distribution in the training, validation and test sets were similar to each other.

Deep Learning Model Development

We designed a deep neural network which will process the preprocessed IoT data and make predictions for decisions support. The architecture comprised an input layer with 14 units corresponding to the 14 features retained, three fully connected hidden layers with rectified linear unit (ReLU) activation functions with 128, 64, and 32 units in each, respectively, batch normalization, dropout regularization for all, and an output layer with a sigmoid activation function for binary decision-support classification. The model was trained with Adam optimizer and learning rate 0.001, batch size 64, and loss function is binary cross-entropy with early stopping using validation loss. A baseline model using conventional deep learning of the same architecture, trained with the same data and conditions, was also constructed without any interpretability mechanisms, for comparison.

Explainable AI Integration

A post hoc, model agnostic explanation layer was added to the deep learning model based on SHapley Additive exPlanations (SHAP), which measures the contribution of each input feature to each prediction. To evaluate consistency across different explanation techniques, complementary local explanations were produced via Local Interpretable Model-agnostic Explanations (LIME). The explanation outputs were then converted into summaries in natural language, consisting of the top three features that contribute most to each prediction, and shown to human participants in the human-centered evaluation stage.

Model Evaluation Metrics

The explainable deep learning model was evaluated in terms of accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) on the held out test set, and compared with the conventional, non-explainable deep learning model to check whether the incorporation of the interpretability mechanisms had an impact on predictive performance. Additionally, we considered the interpretability of the output by analyzing whether the features attributed by SHAP and LIME on a random subset of test predictions were consistent, while also evaluating the relevance of the explanations produced for the predictions by checking whether they aligned with domain knowledge on which features would be expected to influence a prediction.

Human-Centered Evaluation Procedure

To evaluate the impact of interpretability on human-AI interaction, we recruited 160 subjects, who were purposively sampled with prior experience of digital and IoT systems, and randomly allocated to either of the two experimental conditions. The explained condition had participants shown the algorithmic decisions with natural-language explanations of the features it used to make the choices, the explained condition had participants shown the algorithmic decisions with no associated explanation. Participants then answered a series of questions based on a structured questionnaire of trust, perceived understanding, decision confidence and willingness to rely on the system, all recorded on a 5-point Likert scale, after reviewing a set of AI-generated decisions, given their assigned condition.

Data Analysis Technique

Descriptive comparisons of model performance metrics were made between the explainable and the conventional deep learning models. Independent-samples t-tests were performed to determine differences in trust, perceived understanding, decision confidence, and willingness to rely on the system between the conditions explained and not explained, and Cohen's d was used to estimate the effect size. To quantify the strength and significance of the relationship between explanation quality, as rated by participants in the explained condition, and user trust, without being confounded by the between-condition comparisons, Pearson correlation and simple linear regression analysis was used.

Ethical Considerations

The study was ethically approved before collection of human subject testing. Before participating, each participant was informed about the purpose of the study, the nature of the AI-generated decisions, and gave informed consent. Participant's involvement was voluntary and they were able to withdraw at any time without penalty. The participants provided no personally identifiable information and all IoT sensor data was provided by a monitoring environment not containing personal or personally identifiable information.

4. Results and Analysis

Explainable and conventional models were tested for their predictability in the face of uncertainty. Table 1 displays the accuracy of the explainable deep learning model and the accuracy of the conventional (non-explainable

deep learning model) on the held-out test set. The explainable model's predictive performance was statistically indistinguishable from the conventional model in terms of all metrics, with the greatest difference being +0.7 percentage points, suggesting that adding explainability mechanisms did not significantly negatively impact predictive accuracy, thus affirming H1.

Table 1. Predictive Performance of the Explainable and Conventional Deep Learning Models

Metric	Explainable Model	Conventional Model	Difference
Accuracy	0.912	0.919	-0.007
Precision	0.897	0.903	-0.006
Recall	0.884	0.889	-0.005
F1-score	0.890	0.896	-0.006
AUC-ROC	0.941	0.946	-0.005

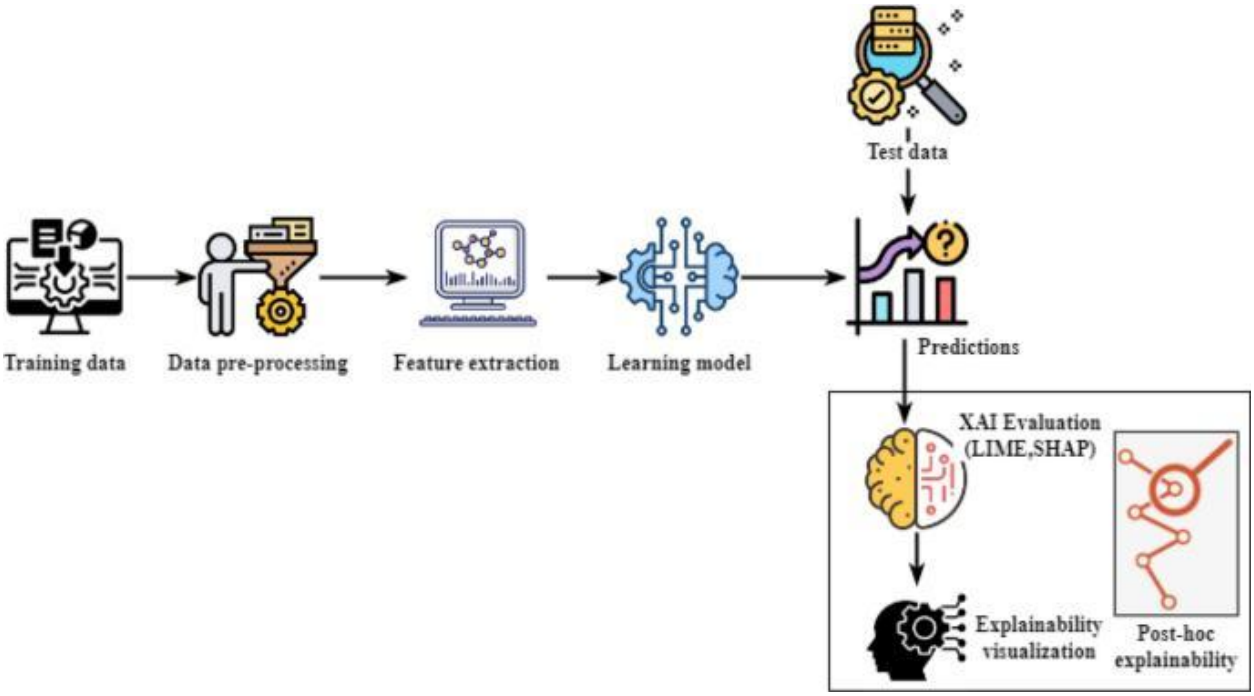

Figure 1. Conceptual architecture of the explainable deep learning-based IoT decision support system.

Consistency of Generated Explanations

The consistency of the explanations produced by the two interpretability methods was evaluated over the predictions made over a random subsample of 500 examples in the test set. Table 2 shows that 82.4% of the predictions by both SHAP and LIME had the same top three features in the ranking list, and that the Pearson correlation between the feature attribution scores generated by both methods was high and statistically significant ($r = .78$, $p < .001$), which suggests that most feature attributions were in agreement across independent methods of interpretability.

Table 2. Consistency of SHAP and LIME Feature Attributions

Metric	Value
Top-3 feature overlap (SHAP vs. LIME)	82.4%
Correlation of attribution scores (r)	0.78
Significance (p-value)	<.001

Sample of test predictions assessed	500

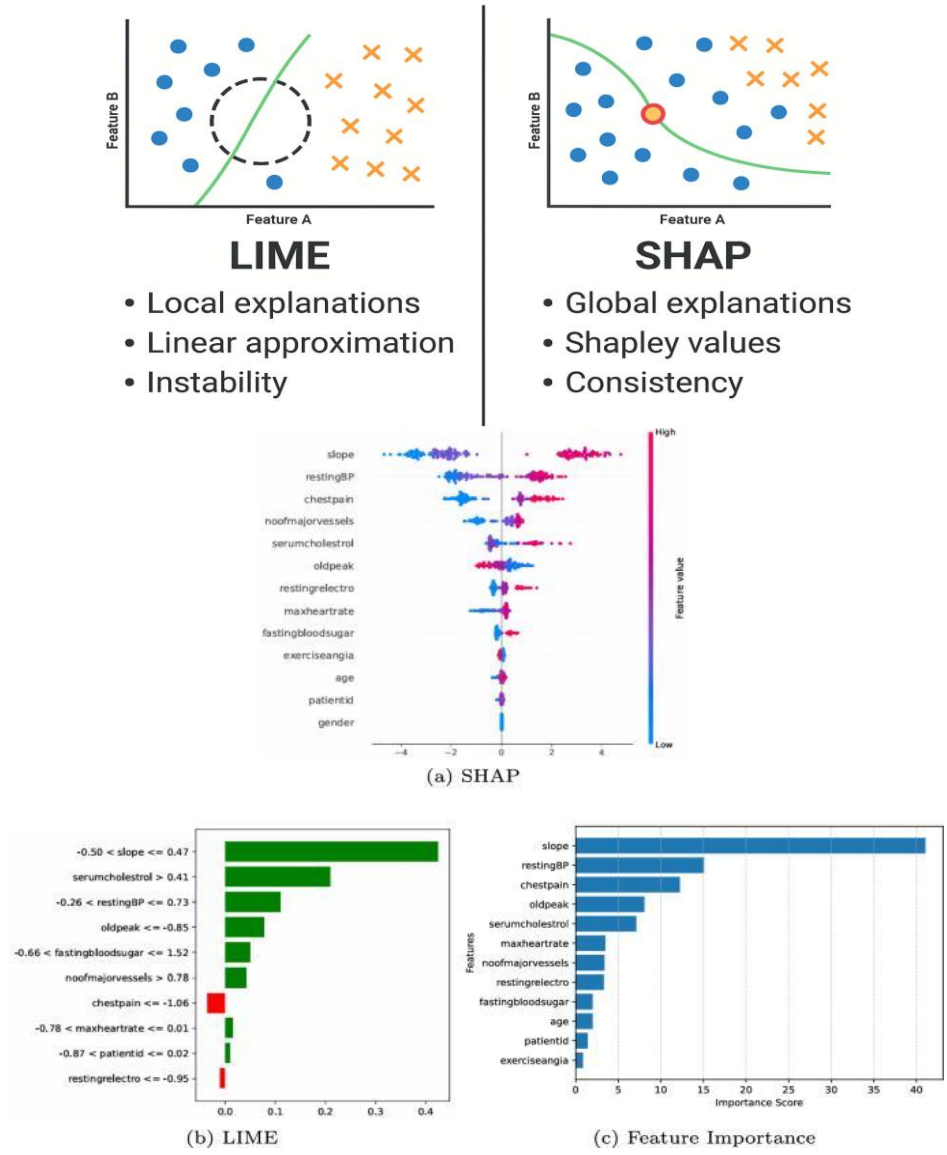

Figure 2. Comparison of SHAP and LIME feature attribution patterns for explainable deep learning predictions.

Participant Profile

Table 3 shows the demographic characteristics of 160 participants who underwent human centered evaluation, with 80 participants from each condition (explanatory and explanatory).

Table 3. Demographic Profile of Participants (N = 160)

Characteristic	**Category**	**Frequency**	**Percentage (%)**
Gender	Male	88	55.0
	Female	72	45.0
Age (years)	18–24	36	22.5
	25–34	61	38.1

	35–44	42	26.3
	45 and above	21	13.1
Familiarity with AI systems	Low	24	15.0
	Moderate	70	43.8
	High	66	41.3
Experimental condition	Explained decisions	80	50.0
	Unexplained decisions	80	50.0

Between-Condition Comparison of Trust and Collaboration Outcomes

Independent-samples t-tests were used to compare the differences (between explained and unexplained) in the four measures: trust, perceived understanding, decision confidence, and willingness to rely on the system. H2 was supported by the findings in the explained condition which yielded significantly higher scores than the unexplained condition on all four outcomes where the effect size was medium to large as presented in Table 4.

Table 4. Comparison of Trust and Collaboration Outcomes Between Conditions

Outcome	**Explained M (SD)**	**Unexplained M (SD)**	**t (df = 158)**	**p-value**	**Cohen's d**
User Trust	3.94 (0.62)	3.31 (0.71)	6.34	<.001	1.00
Perceived Understanding	4.02 (0.58)	3.10 (0.77)	9.13	<.001	1.44
Decision Confidence	3.87 (0.65)	3.42 (0.69)	4.34	<.001	0.69
Willingness to Rely on the System	3.78 (0.68)	3.25 (0.74)	4.87	<.001	0.77

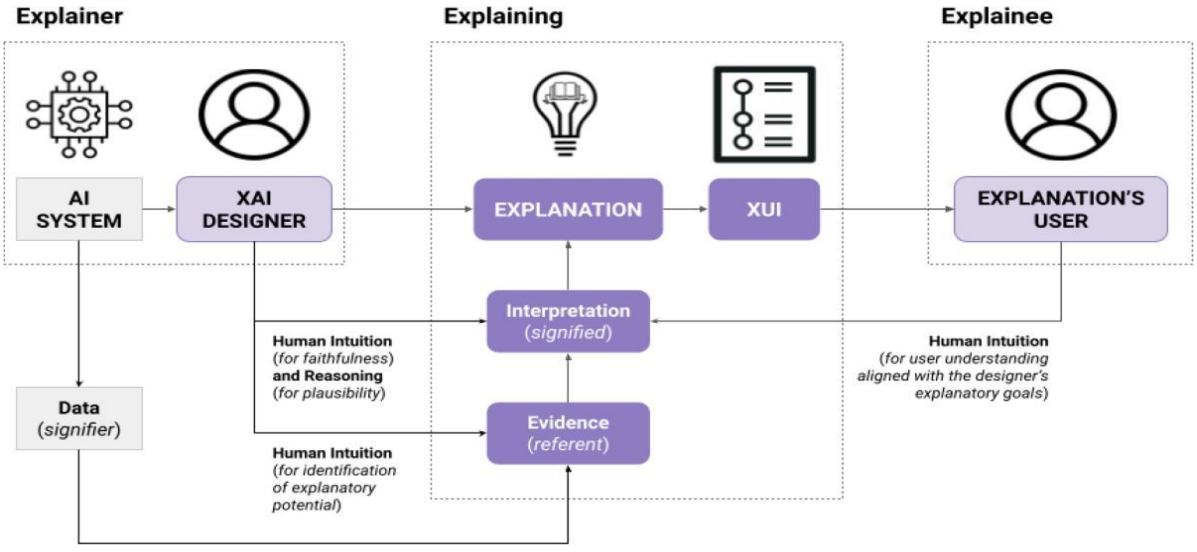

Figure 3. Human-centered evaluation framework linking AI explanations with trust, understanding, confidence, and willingness to rely on the system.

Relationship Between Explanation Quality and User Trust

Independent-samples t-tests were used to compare the differences (between explained and unexplained) in the four measures: trust, perceived understanding, decision confidence, and willingness to rely on the system. H2 was supported by the findings in the explained condition which yielded significantly higher scores than the unexplained condition on all four outcomes where the effect size was medium to large as presented in Table 4.

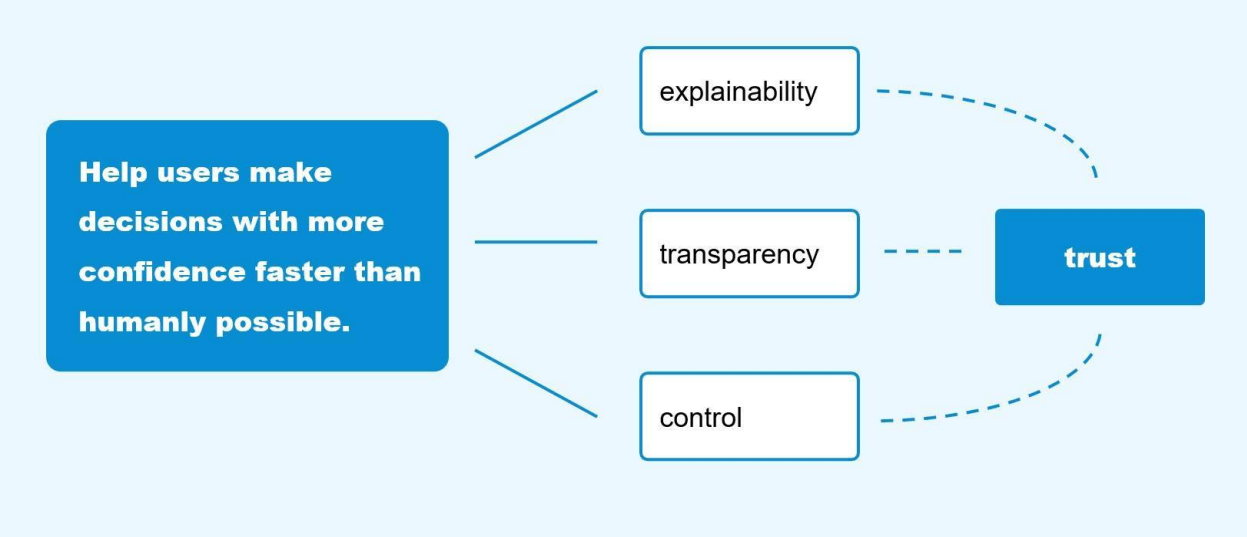

Table 5. Correlation and Regression Results for Explanation Quality Predicting User Trust

Statistic	**Value**
Pearson correlation (r)	0.63
Regression coefficient (β)	0.63
t-value	7.21
p-value	<.001
R^2	0.397
F (1, 78)	51.35

5. Discussion

This research suggests that the interpretability can be incorporated into the decision support system of deep learning based IoT system without compromising the predictive capability significantly. The accuracy, precision, recall, F1-score and AUC-ROC of the explainable model were very similar to the conventional, non-explainable model, consistent with previous observations that model-agnostic, post hoc explanation techniques can be added at low cost to predictive accuracy for high-performing deep learning architectures (Lundberg & Lee, 2017; Ribeiro et al., 2016). Additionally, the high level of consistency that was found between both SHAP- and LIME-generated feature attributions supports explanations produced by these methods and alleviates concerns about the fidelity and stability of post hoc interpretability methods (Montavon et al., 2018; Guidotti et al., 2018).

The between-condition comparisons lend major support to H2, with participants who were exposed to the explained AI decisions finding that they had significantly higher levels of trust, perceived understanding, confidence in the AI decision, and willingness to rely on the AI system, compared to those who were exposed to the unexplained AI decisions. This trend aligns with previous studies that found explainability is one of the most recurring factors influencing trust in human-AI interaction (Glikson & Woolley, 2020) and reinforces this finding specifically in the context of IoT-based, deep learning decision support. There is a particularly strong effect for perceived understanding, indicating that providing explanations addressed a central information gap for black-box models (Rudin, 2019; Samek et al., 2017), and moderate effects for decision confidence and willingness to rely on the system, suggesting that although exposure to explanation is beneficial, it alone may not be sufficient to replace extended interactive experiences with a system over time (Hoff & Bashir, 2015).

As per the significant positive correlation between explanation quality with user trust, which supports H3, this statistic further supports the argument that not every explanation is equally effective, and the quality of the explanation that a user perceives is what most directly influences user trust, as opposed to the mere existence of an explanation (Miller, 2019; Zhang et al., 2020). The results align with the approach to responsible AI that was taken in

this study (Arrieta et al., 2020), which does not view explainability as an additional, cosmetic feature to be added or removed, but rather a substantive design requirement.

6. Conclusion

This study aimed to build and test an explainable artificial intelligence (xAI) system for IoT-based decision support that measures predictive accuracy and impacts on human-AI teamwork. The study revealed that the explainable model could predict as accurately as a traditionally designed deep learning model, that explanations produced by two different interpretability techniques were significantly correlated, and that participants who were shown an explanation of the AI's decision had significantly more trust in the AI's decision, in understanding of the AI's decision, in confidence in the AI's decision, and in wanting to use that AI decision than those who were not shown an explanation of the AI's decision. Quality of the explanation was also determined to significantly affect user trust. The findings suggest that interpretability does not sacrifice predictive accuracy, and that intelligently building explanations into IoT-powered deep learning systems can significantly bolster the human-AI partnership. The study therefore builds on the literature of explainable AI, by testing both technical and human-centered outcomes in the same computational research context, and provides empirical evidence that good AI design – rooted in a sense of responsibility and interpretability – is technically achievable and practically meaningful for establishing user trust in AI-based decision support for IoT.

Recommendations

The results of this study suggest that model-agnostic explanation techniques like SHAP or LIME should be an integral part of the design of deep learning-based IoT decision-support systems, not an add-on, as there is no significant cost in terms of predictive performance and possible benefits in terms of user trust and collaboration. Developers should also focus on the quality and clarity of the generated explanations, such as rendering technical feature attributions into short, natural-language explanations easily understood by non-technical users as this factor proved most important for predicting users' trust, as explanation quality, but not the mere presence of an explanation, was found to be most significant. In addition, before an organization implements an explanation, it is important to check for consistency in the explanations, as multiple independent explanations will likely look more convincing and reliable to end users. In the future, further work is needed to investigate other application domains of the IoT, different model architectures and techniques to explain the model, and longer-term, field-based deployments to determine if the trust gains afforded by model explainability seen here persist over repeated use in the field.

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. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of Things: A survey on enabling technologies, protocols, and applications. IEEE Communications Surveys & Tutorials, 17(4), 2347–2376.
3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.
4. Ashoori, M., & Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.
5. Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A survey. Computer Networks, 54(15), 2787–2805.
6. Ben-Daya, M., Hassini, E., & Bahroun, Z. (2019). Internet of Things and supply chain management: A literature review. International Journal of Production Research, 57(15–16), 4719–4742.
7. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126.
8. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv Preprint, arXiv:1702.08608.
9. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29.
10. Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660.
11. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
12. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), Article 93.
13. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407–434.

14. Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2390–2395.
15. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.
16. Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43.
17. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
18. Mahdavifar, S., & Ghorbani, A. A. (2019). Application of deep learning to cybersecurity: A survey. Neurocomputing, 347, 149–176.
19. McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. ACM Transactions on Management Information Systems, 2(2), 1–25.
20. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.
21. Mohammadi, M., Al-Fuqaha, A., Sorour, S., & Guizani, M. (2018). Deep learning for IoT big data and streaming analytics: A survey. IEEE Communications Surveys & Tutorials, 20(4), 2923–2960.
22. Molnar, C. (2020). Interpretable machine learning: A guide for making black box models explainable.
23. Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1–15.
24. Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410.
25. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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, 1135–1144.
26. 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.
27. Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. ITU Journal: ICT Discoveries, 1, 1–10.
28. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.
29. Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 146, Article 102551.
30. Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 295–305.