Explainable AI-Driven Threat Detection and Response for Industrial IoT

Shokooh Khandan*, Deniz Beyazgul**, Olamide Jogunola*, Yakubu Tsado*, Tooska Dargahi*
Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK.
Email: *{s.khandan, o.jogunola, y.tsado, t.dargahi}@mmu.ac.uk; **deniz.beyazgul@stu.mmu.ac.uk

***Abstract*—The growing complexity and volume of cyber attacks to Cyber-Physical Systems (CPS) and Industrial Internet of Things (IIoT) have outpaced traditional detection methods, requiring more intelligent and explainable security solutions. While Artificial Intelligence (AI)-based anomaly detection solutions have been proposed in the literature, they either focus on a single type of attack, or their decisions are restricted based on a single dataset, or they lack transparency. To address these challenges, this paper presents an explainable attack detection framework for IIoT, combining advanced machine learning (ML) models and AI-driven interpretability. The framework employs mid-level and late data fusion techniques on two IIoT datasets, using Autoencoders (AE) and Manifold Alignment (MA) techniques to generate a unified feature space. A Random Forest (RF) classifier is trained on the fused dataset to detect four attack types, achieving a 97% accuracy. The model's decision-making is made transparent through Explainable AI (XAI) tools, providing both global and local interpretability. Furthermore, a Large Language Model (LLM)-powered AI assistant is developed to provide automated, context-aware mitigation strategies based on MITRE D3FEND framework. This integrated approach enhances the detection, interpretability, and response to threats in IIoT environments, promoting greater trust and operational resilience.**

***Index Terms*—Explainable AI, Cyber-Physical Systems, Industrial IoT, Cyber Attack Detection**

I. Introduction

Cyber-physical systems (CPS) are central to industrial internet of things (IIoT) deployments, linking sensors, actuators, and control processes to enable automated industrial operations. Such systems often underlie critical sectors, where any disruption can have serious consequences. The interconnected IIoT devices increase the attack surface and render conventional defences inadequate [1]. As IIoT adoption expands, threats such as malware, unauthorised access, and denial-of-service attacks have become more prevalent, often overwhelming traditional defence [2]. The UK's National Cyber Security Centre managed 430 cyber incidents between September 2023 and August 2024, 89 of which were deemed nationally significant. These incidents impacted various sectors, including healthcare and public services [3]. Therefore, effective attack detection systems are essential for the timely identification and mitigation of anomalies in IIoT networks [1], [2]. Artificial intelligence (AI), particularly machine learning (ML), has emerged as a promising solution for attack detection in CPS, since the conventional defence against these threats are inadequate [4]. AI-based solutions provide data-driven anomaly detection and adaptability to novel threats. For example, ML-based frameworks demonstrate enhanced detection accuracy in heterogeneous IoT environments [4]. However, IIoT generates diverse multidomain datasets, including network traffic, sensor signals, and system logs, necessitating fusion and alignment techniques to improve attack detection. Manifold alignment methods align data distributions across domains by preserving intrinsic structures, enabling effective comparison and combination of datasets [5]. Researchers suggest that models trained on a combined dataset provide better generalisation to enhance detection accuracy and resilience [6]. Babayigit and Abubaker [5] fused three IIoT datasets and used a hybrid deep learning model with multiple-domain and transfer learning. However, including the Edge-IIoTSet adds complexity and imbalanced data challenges, reducing accuracy due to diverse attack types and complicating latent space interpretation. To address this complexity, in this paper, we propose a framework which employs an Autoencoder (AE) and Locally Linear Embedding (LLE) methods for manifold alignment to detect four attack categories including Distributed Denial of Service (DDoS)/DoS, code injection, malware and reconnaissance, achieving a 97% accuracy. The main contributions of the paper include:

- We created a fused dataset for investigating attack patterns in IIoT using two IIoT datsets in the literature (i.e. X-IIoTID and WUSTL-IIoT-2021). The new combined dataset has a latent space of 30 dimensions, which preserves more nuanced and comprehensive information compared to the smaller latent dimension (around 10 dimensions) utilised in [5].
- The proposed framework uses Explainable AI (XAI) techniques, including local and global explaination techniques, to interpret and justify the decisions that our AI models have made to identify the four types of attacks.
- We have integrated a Large Language Model (LLM)-powered AI assistant within our approach to provide automated, context-aware mitigation recommendations for each of the identified attack types, based on the MITRE D3FEND [7] guidelines.

In the remainder of this paper, Section II reviews the related work. Data pre-processing methodology is discussed in

This research was supported by the "VISTA – Visual Intelligent System for Threat Analysis" project funded by The Alan Turing Institute under Grant agreement number G2040.

Section III. The proposed framework and experimental analysis are presented in Section IV, while Section V presents the LLM-based attack mitigation recommender system. Section VI summarises the findings and outlines future research directions.

II. Literature Review

Recent research studies on IIoT attack detection have focused on data pre-processing to boost model accuracy. In [8], the authors utilise the synthetic minority over-sampling technique (SMOTE) to address class imbalance. They trained multiple centralised ML models for binary and multi-class classification, achieving high accuracy in binary tasks. However, while synthetic oversampling methods like SMOTE-ENN help balance class distributions, they risk adding artificial or redundant samples that may not reflect real-world attack patterns. Authors in [8], [10] compare Centralised Deep Learning (CDL) with Federated Deep Learning (FDL) for IIoT context. They train separate DL models on each dataset for both binary and multi-class classification without merging the datasets. Their findings show that FDL consistently outperforms CDL in terms of accuracy, while maintaining data privacy. Bahadoripour et al. [9] propose a deep federated multi-modal framework for cyber-attack detection in ICS, integrating representation learning, domain adaptation, and FL. However, the model's reliance on a well-suited public dataset for domain adaptation presents limitations, if the dataset fails to represent diverse client distributions, model accuracy and generalisability suffer.

To improve generalisability of these models, researchers propose combining several datasets using fusion methods to enhance detection accuracy of the models. In [5], the authors train CNN-GRU, a hybrid DL model with a Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU), CNN-GRU classification model on an IIoT combined dataset. Both a binary model and a multi-classification model (6 attack types) are proposed. Combining datasets improves attack detection rate, however, the developed intrusion detection systems (IDS) often function as ambiguous "black boxes" limiting their effectiveness in critical infrastructure, including IIoT, where transparency and interpretability are crucial. Therefore, very recently, researchers have proposed the integration of XAI into IDS [11]. Authors propose xIIRS framework that employs explainable deep learning for intrusion response in industrial settings. By calculating feature importance scores, it aids in understanding AI decisions, thereby supporting trustworthiness and compliance. XAI has been used in other CPS scenarios, such as industrial sensor networks and autonomous traffic sign recognition [12], as well as resource-constraint environments [13]. The reviewed literature typically focus on enhancing a single aspect, such as pre-processing improvements, dataset integration, detection of a single attack. There are also a limited number of papers on XAI adoption in CPS applications. Instead, our proposed approach consolidates all these aspects into a unified framework that enhances threat detection, interpretability, and response capabilities within IIoT context. Our proposed framework prioritises AI explainability and develops a multi-classification RF model to detect several attack categories. Explainability is critical in practical deployments, especially in IIoT environments, where understanding model predictions is crucial for decision-making and security management. Table I presents a comparative analysis of our proposed approach with the literature that we covered in this section.

III. Data pre-processing Methodology

The following subsections explain the data pre-processing and Exploratory Data Analysis (EDA) on each dataset that are performed before feeding them into the model for training.

A. Dataset Features

1) X-IIoT Dataset: The X-IIoTID dataset is created by researchers at the University of New South Wales. The dataset is designed to be connectivity and device-agnostic, which makes it suitable for the heterogeneous nature of IIoT systems. It includes normal and malicious records (roughly 421,417 normal and 399,417 malicious) with 66 features, covering network traffic, device resources, and logs. It supports several protocols and includes diverse attack scenarios such as reconnaissance, ransomware, and DDoS.

Our target variable in the dataset is Attack_Type. The columns Date, Timestamp, Scr_IP, Scr_port, Des_IP, Des_port,

TABLE I
Review of the Related work

Ref	Dataset	Models	Combined Datasets	XAI	LLM
[5]	Edge-IIoTSet, WUSTL-IIoT-2021, X-IIOTID	CNN-GRU	Yes	No	No
[6]	Real time PLC data aquisition	Decision Fusion (Rule based, DT, SVM, LSTM, XGBoots)	No	No	No
[9]	Water treatment and gas pipeline	Deep Federated multi-modal model	Yes	Global SHAP	No
[10]	X-IIoTID, Edge- IIoTset, and WUSTL-IoT-202	Centralised and Federated Deep Learning	No	No	No
[11]	ON-IoT and Gas Pipeline	Long short-term memory (LSTM), AE	No	Global SHAP, LIME, LEMNA	No
[12]	CPS system simulations	CNN+XAI	No	Global SHAP	No
[13]	Edge-IIoTset, UKM-IDS20, CTU-13, and NSL-KDD	Lightweight XAI Network Security framework(LENS-XAI)	No	Variable attribution-based	No
Our Approach	WUSTL-IIoT-2021, X-IIOTID	Random Forest (RF)	Yes	Global SHAP, Local SHAP, LIME, Feature Importance	Yes

class1, and class3 are dropped prior to EDA, resulting in a final dataset with 820,834 rows and 60 columns. Figure 1 shows the distribution of the attack types in this dataset (where "normal" means benign).

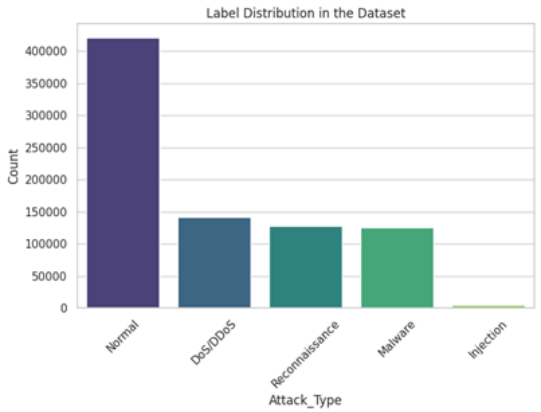

Fig. 1. Distribution of `Attack_Type` in the X-IIoT dataset.

2) WUSTL-IIoT-2021 Dataset: The Wustl_iiot_2021 dataset is developed by Washington University in St. Louis for cyber security research in IIoT environments. It includes network traffic data collected from a testbed emulating real-world industrial systems, featuring both normal operations and various cyber attacks. The target variable in this dataset is Attack_Type. The columns StartTime, LastTime, SrcAddr, DstAddr, Sport, Dport, and Target are dropped prior to EDA, resulting in a final dataset with 1,194,464 rows and 49 columns.

B. Dataset Labels

A key challenge in multi-source data integration is the inconsistency in data labelling, as each dataset originally defines attack types with different granularity. To address this, we restructure the original attack/benign labels into five consolidated categories to improves generalisability, enhances model performance, and simplifies interpretation.

- DoS/DDoS (e.g., RDoS, DoS),
- Reconnaissance (e.g., reconnaissance, Reconn),
- Malware (e.g., crypto-ransomware, weaponisation, exfiltration, command & control, backdoor, tampering),
- Injection (e.g., command injection, exploitation),
- Normal.

IV. Proposed Framework and Analysis

The first steps in our framework include data pre-processing, feature transformation, and fusion techniques on the two IIoT datasets. We then selected RF technique due to its superior explainability compared to the other ML models, and trained an RF multi-classification model to predict the four attack types in Section III-B. Then XAI techniques are applied to the model to provide explainability and increase the trustworthiness of the proposed approach. Finally, an AI-assistant has been trained to provide automated, context-aware response recommendations for high-priority alerts. Figure 2 presents each component of the proposed framework, as explained below.

1) Data Pre-processing: Before feeding the data into a model, it is crucial to clean and normalise it. This ensures that all features are on a comparable scale, which is essential for effective learning.

2) Mid Fusion via Feature Transformation: This step identifies the 12 common features shared across both datasets. These features provide a consistent basis for aligning the data.

3) AE Training: An Autoencoder (AE) is trained on the mapped and pre-processed features to create a compressed, yet meaningful representation of the data. This latent space helps reduce noise and dimensionality, retaining only the most informative characteristics.

4) Manifold Alignment: After transforming both datasets using the AE, a technique called Modified Locally Linear Embedding (M-LLE) is used to align them. M-LLE seeks a lower-dimensional projection of the data which preserves distances within local neighbourhoods. This preserves the internal structure of each dataset while projecting them into a shared feature space.

5) Combined Dataset Output: The latent space representation from the encoder is extracted and used as the transformed feature set for downstream tasks. By combining AE-based feature learning with MA, mid fusion is effectively achieved, allowing for robust cross-dataset analysis while preserving the integrity of both datasets. During the late fusion stage, decision-level fusion is utilised by training separate models on the two datasets and integrating their outputs at the decision stage. After applying the data fusion process, the combined

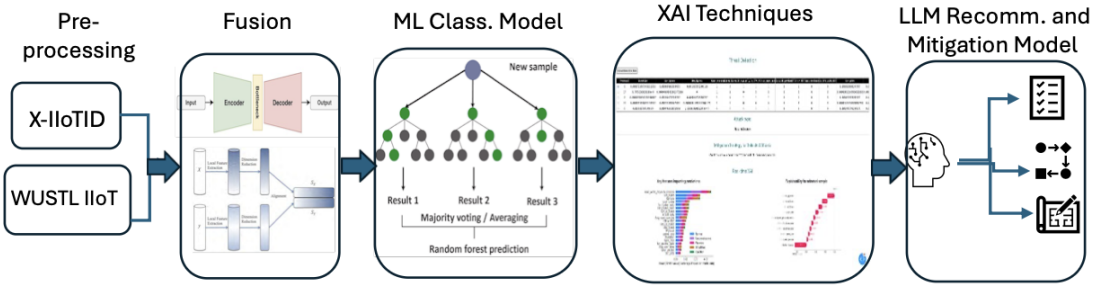

Fig. 2. The proposed attack detection and recommendation framework.

dataset is transformed into a latent space representation with 32 features, effectively capturing the most informative characteristics from both datasets.

6) Model Development: An RF classifier model is trained on the fused dataset. It is selected to ensure reliable detection and ease of understanding in real-world deployment scenarios.

7) Explainable AI (XAI): Explainability refers to the degree to which the model's decisions can be understood [14]. Complex models also known as black-box models, such as deep neural networks (DNNs) and RF, are often criticised for their lack of transparency [14]. Hence, XAI techniques aim to overcome these limitations. Local explanations, such as local SHapley Additive explanation (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), offer insight into individual predictions and interpret models locally around each prediction [15]. This helps users understand specific outcomes while using complex models, which are otherwise unclear. Global explanations, such as feature importance and global SHAP, are other approaches to explainability. Global explanations provide insight into how individual features impact the model behaviour on the historical data (training datasets). SHAP values, originating from cooperative game theory, offer a unified approach to explain the predictions of the models [16].

A. Experimental Performance Analysis

The developed RF model is evaluated considering the following performance metrics: accuracy, precision, recall, and F1-score metrics. It performs exceptionally well, achieving 97% accuracy in multiple types of attacks, as shown in Table II.

TABLE II
RF MULTI-CLASS MODEL PERFORMANCE METRICS.

Class	Accuracy	Precision	Recall	F1-Score
DoS/DDoS	0.97	0.98	0.99	0.99
Injection	0.97	0.99	0.78	0.55
Malware	0.97	0.90	0.88	0.88
Reconnaissance	0.97	0.98	0.99	0.99
Normal	0.97	0.97	0.87	0.87

Table III highlights the key differences between our approach and the baseline [5]. We avoid synthetic oversampling to reduce the risk of introducing unrealistic data that may impair model generalisation to real attack scenarios. We only select homogeneous datasets. Retaining a higher latent space dimension preserves nuanced information, enables the model to capture diverse and subtle attack patterns more effectively. We prioritise interpretability and data authenticity to support more reliable and explainable threat detection outcomes.

Figure 3 presents the feature importance plot, showing a few top-ranking features, 'Service', 'read_write_physical.process', and 'Av_num_cswch/s', highlighting their critical roles in identifying anomalous or operational states in industrial systems. These features reflect service-level characteristics, low-level physical process interactions, and CPU context-switching behaviour, which are all essential indicators of system health and potential security threats. Other highly ranked features, such as 'Duration' further highlights the importance of temporal

TABLE III
COMPARISON OF OUR APPROACH WITH THE BASELINE APPROACH [5]

Aspect	Baseline Approach [5]	Our Approach
Datasets	Edge-IIoTSet, WUSTL-IIoT-2021, X-IIOTID	X-IIoTID, WUSTL-IIoT-2021
Dataset similarity	Combined datasets with significant heterogeneity (Edge-IIoTSet)	Selected more homogeneous datasets
Class imbalance handling	Used SMOTE-ENN	Avoided synthetic data
Latent space dimensions	Reduced to 10 dimensions	Retained higher latent space of 30 dimensions
Focus	Generalisation across diverse datasets	Interpretability and coherence within similar datasets
Accuracy, Recall, Precision	0.97, 0.97, 0.97 (CNN-GRU model)	0.97, 0.90, 0.96 (RF model)

and traffic-related metrics. The long-tail feature importance pattern indicates that a few key features drive most of the model's predictive power, while many others add only minor improvements, which aids efficient feature selection and improves model interpretability in real-world IIoT applications

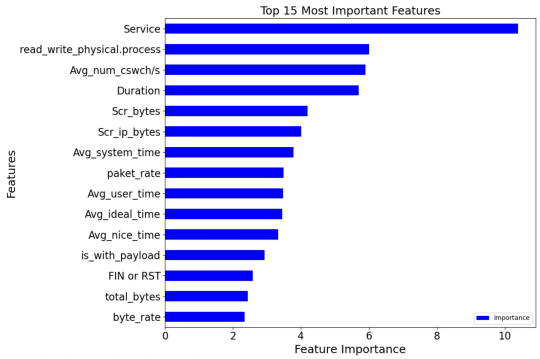

Fig. 3. Feature importance for IIoT dataset, revealing how each feature contributes to the model's prediction.

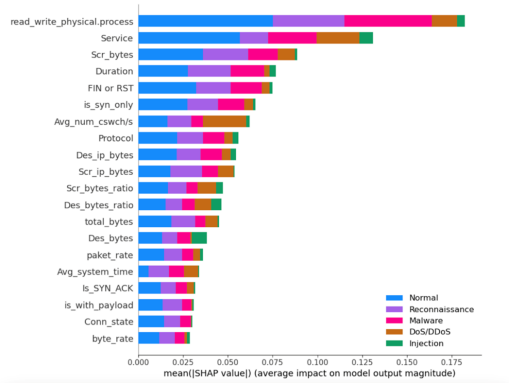

Fig. 4. Global SHAP-based model interpretability for IIoT dataset, revealing how each feature changes the model's prediction to classify cyber threats.

Figure 4 displays a global SHAP summary highlighting the top features influencing the model's network behaviour classifications across attack types. The feature

'read_write_physical.process' shows the highest overall contribution across nearly all classes, particularly in distinguishing malware and reconnaissance attacks. Other significant features include 'Service', 'Scr_bytes', and 'Duration', which play varying roles depending on the attack type. This visualisation demonstrates not only which features are most important overall but also how their influence differs across specific cyber threat categories in the IIoT context.

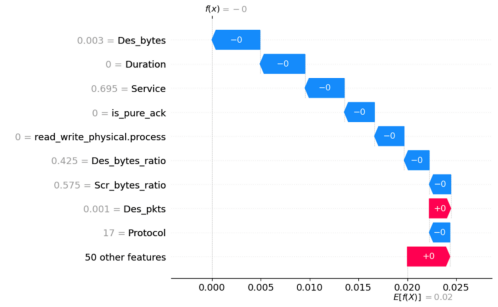

Fig. 5. Local SHAP explanation for a malware attack instance in the IIoT dataset.

Figure 5 shows how feature SHAP values led the model to predict a Malware Attack. Most features in this case, such as 'Service', 'Duration', 'FIN' or 'RST', and 'Scr_bytes_ratio', have negative SHAP values (blue bars), meaning they pull the prediction away from "non-attack" and help confirm the Malware classification. One feature, 'read_write_physical.process', had a slight positive impact (red bar), temporarily increasing the likelihood of a non-attack classification, but not enough to change the outcome. Overall, the prediction is shifted from a base value of 0.18 to 0, strongly confirming the presence of Malware activity. This plot provides insight into which sensor-level indicators were most influential in detecting the threat.

Figure 6 presents the local SHAP explanation for a Normal instance. The plot shows how each feature's SHAP value supports the model's normal classification. Most features, such as 'Service', 'FIN or RST', 'is_syn_only', 'Protocol', and 'Duration', have positive SHAP values (red bars), meaning they push the prediction toward the "normal" class. One feature, 'read_write_physical.process', had a negative SHAP value, slightly pulling the model toward an attack prediction, but not enough to outweigh the strong normal indicators. The prediction starts at a base value of 0.579 and is confidently shifts to 1.0, confirming that the model strongly recognises this behaviour as benign. This plot demonstrates how the model uses a combination of subtle traffic and protocol features to confidently detect non-malicious activity.

Figure 7 presents the LIME analysis for a specific instance classified as a DoS/DDoS attack. The bar chart highlights the most influential features that contribute to the prediction. Among the top contributors to the DoS/DDoS decision were 'Scr_bytes_ratio' 0.67, Avg_user_time 0.22, and Avg_num_cswch/s, each providing positive support for the classification. The feature values on the right show the actual observed inputs for the instance, such as a high Scr_bytes_ratio

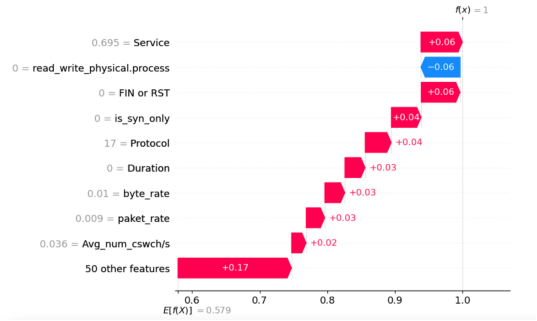

Fig. 6. Local SHAP explanation for a normal (non-attack) instance in the IIoT dataset.

1.00, elevates 'Avg_user_time' 0.69, and a high 'Service' 0.89, all of which align with typical behaviours of DoS/DDoS activity, such as high source byte ratios and CPU load. Conversely, features like 'Des_bytes_ratio' 0.33 and FIN or RST 0.00 contribute marginally but still play a role in the local decision.

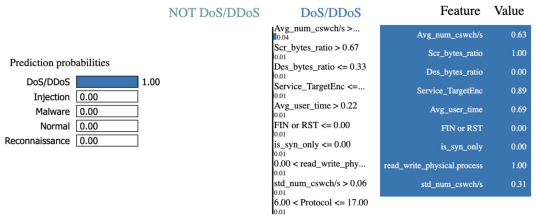

Fig. 7. LIME for a Dos/DDoS attack instance in the IIoT dataset.

Figure 8 presents the LIME interpretability analysis for a correctly classified Normal instance. The features contributing most to the classification are visualised, with all top contributors supporting the "Normal" label. Notably, features such as Is_SYN_ACK, dTtl, DstLoss, Scr_pkts, and Scr_bytes, all with values equal to 0 or low thresholds, strongly reinforce the model's decision. These conditions typically reflect the absence of abnormal traffic patterns, packet loss, or suspicious payload behaviour. Additionally, attributes like TotAppByte, SrcLoad, and total_bytes being 0.00 further indicate minimal network activity, characteristic of a benign IIoT environment. Only sTos had a minimal neutral or slightly contrary influence. This explanation confirms that the model bases its decision on a coherent and logical combination of low-activity indicators, which aligns with expected behaviour for non-malicious traffic in IIoT systems.

V. AI-ASSISTED ATTACK MITIGATION RECOMMENDATION

This section introduces an AI-assistant powered by LLM that delivers automated, tailored recommendations for specific cyber attacks in IIoT environments. The LLM is trained on a structured PDF file compiled from best-practice mitigation strategies based on the MITRE D3FEND framework, covering topics such as Remote Denial of Service (RDOS), Fake Notifications, Brute Force Attacks, Man-in-the-Middle (MitM)

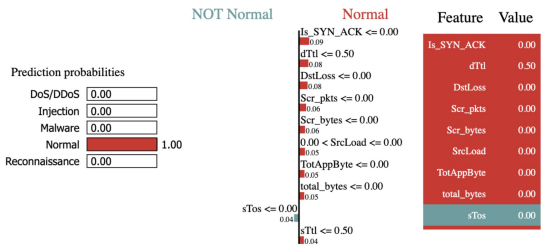

Fig. 8. LIME explanation for a normal instance in the IIoT dataset.

Attacks, Data Exfiltration, Ransomware and Intrusion Detection and Prevention Systems (IDS/IPS) [7]. The system uses LangChain, ChromaDB, and the Flan-T5 model. The PDF file is processed into overlapping text chunks, embedded using all-MiniLM-L6-v2, and stored in ChromaDB for retrieval. When queried, the system retrieves relevant chunks through similarity search, which Flan-T5 uses to generate structured, context-aware answers. Figure 9 presents an excerpt of a question and answer from the developed LLM-based AI-assistant. A sample interaction shows actionable guidance against DoS/DDoS attacks, like SYN cookies and disaster recovery planning. The assistant can also be retrained on other documents, including internal security policies or standards such as ISO 27001 and NIST.

A qualitative human evaluation confirmed the chatbot's responses were accurate, relevant, and fluent, demonstrating the approach's effectiveness in supporting on-demand cybersecurity recommendations.

Q1: *How can we mitigate RDOS (Remote Denial of Service)?*

A1:

- *Implement rate limiting, traffic filtering, and load balancing.*
- *Use DDoS protection services (e.g., Cloudflare, AWS Shield).*
- *Monitor network traffic for sudden surges to detect attacks.*
- *Deploy intrusion detection/prevention systems (IDS/IPS).*

Q2: *How can we mitigate Vul_Scan (Vulnerability Scanning)?*

A2:

- *Obscure system details using security-through-obscurity techniques and firewalls.*
- *Consider using honeypots to detect and analyze scanning behavior*
- *Employ IDS and Firewalls to detect and block unusual scanning behavior.*

Fig. 9. Example questions & answers from the AI-Assistant Chatbot.

VI. Conclusion

In an attempt to improve attack classification and detection accuracy in industrial IoT context, this paper presented a framework that enhances the detection, explainability, and response strategies to cyber attacks. In particular, we selectively combined two datasets using fusion techniques and created a fused data with larger latent dimension for investigating attack detection in IIoT. The fused dataset is then fed into an RF model that achieved an accuracy of 97%, to detect four attack categories. An XAI module is integrated into the framework to provide both global and local interpretability to the decision making of the AI model. We further introduced an LLM-powered AI assistant that delivers automated, context-aware mitigation strategies based on the MITRE D3FEND guidelines. This research reinforces the importance of data fusion in cyber security, demonstrating that integrating multiple datasets through feature transformation and decision aggregation can significantly improve classification accuracy and robustness in IIoT intrusion classification and detection. Future work should focus on class balancing and feature optimisation to enhance recall for under-represented attack types. A comprehensive evaluation of the LLM-based assistant, including user studies has been left as a future work as well.

References

[1] Zhukabayeva, T., Ahmad, Z., Adamova, A., Karabayev, N., & Abdildayeva, A. (2025). An Edge-Computing-Based Integrated Framework for Network Traffic Analysis and Intrusion Detection to Enhance Cyber–Physical System Security in Industrial IoT. Sensors, 25(8), 2395.

[2] Zhukabayeva, T., Zholshiyeva, L., Karabayev, N., Khan, S., & Alnazzawi, N. (2025). Cybersecurity Solutions for Industrial Internet of Things–Edge Computing Integration: Challenges, Threats, and Future Directions. Sensors, 25(1), 213.

[3] Independence (2025). Cyber threat against UK Government severe and advancing quickly, warns watchdog. Available online https://www.independent.co.uk/news/uk/politics/government-national-audit-office-uk-government-geoffrey-cliftonbrown-british-library-b2688062.html?utm. Accessed 28 May, 2025.

[4] Kikissagbe, B. R., & Adda, M. (2024). Machine learning-based intrusion detection methods in IoT systems: A comprehensive review. Electronics, 13(18), 3601.

[5] Babayigit, B., & Abubaker, M. (2024). Towards a generalized hybrid deep learning model with optimized hyperparameters for malicious traffic detection in the Industrial Internet of Things. Engineering Applications of Artificial Intelligence, 128, 107515.

[6] Xue, Y., Pan, J., Geng, Y., Yang, Z., Liu, M., & Deng, R. (2024). Real-Time Intrusion Detection based on Decision Fusion in Industrial Control Systems. IEEE Transactions on Industrial Cyber-Physical Systems.

[7] MITRE D3FEND™: A knowledge graph of cybersecurity countermeasures. Available at: https://d3fend.mitre.org (Accessed: 10 March 2025).

[8] M. Ferrag, O. Friha, D. Hamouda, L. Maglaras and H. Janick (2022) Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications for Centralized and Federated Learning

[9] S. Bahadoripour, H. Karimipour, A. Jahromi, A. Islam (2024) An explainable multi-modal model for advanced cyber-attack detection in industrial control systems

[10] Popoola, S. I., Imoize, A. L., Hammoudeh, M., Adebisi, B., Jogunola, O., & Aibinu, A. M. (2023). Federated deep learning for intrusion detection in consumer-centric internet of things. IEEE Transactions on Consumer Electronics, 70(1), 1610-1622.

[11] Xue, Q., Zhang, Z., Fan, K., & Wang, M. (2025). xIIRS: Industrial Internet Intrusion Response Based on Explainable Deep Learning. Electronics, 14(5), 987.

[12] Taufik, M., Aziz, M. S., & Fitriana, A. (2025). Hybrid Explainable AI (XAI) Framework for Detecting Adversarial Attacks in Cyber-Physical Systems. Journal of Technology Informatics and Engineering, 4(1).

[13] Yagiz, M. A., & Goktas, P. (2025). LENS-XAI: Redefining Lightweight and Explainable Network Security through Knowledge Distillation and Variational Autoencoders for Scalable Intrusion Detection in Cybersecurity. arXiv preprint arXiv:2501.00790.

[14] 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

[15] Lipton, Z. C. (2018). The Mythos of Model Interpretability. Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning

[16] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS)