Explainable AI-Driven Security Framework for Cyber-Physical Production Systems in Industry 4.0: Leveraging Immersive Embedded CIoT

Muhammad Farhan, Muhammad Rehan Naeem, Ahmad Almadhor, Ali Kashif Bashir, Zhu Zhu, Thippa Reddy Gadekallu

***Abstract*—This paper proposes a new security framework of Explainable Artificial Intelligence (XAI) for Cyber-Physical Production Systems (CPPSs) in the Industry 4.0 paradigm. An integral part of the framework is the incorporation of XAI into immersive embedded Consumer Internet of Things (CIoT) systems to promote transparency and interpretability, thereby improving real-time decision-making in AI-driven security mechanisms. The framework leverages SHAP and LIME techniques to provide human operators with clear insights into the AI-based security decision-making process, fostering trust and facilitating effective teaming between humans and AI assets. The proposed solution was validated and tested in an innovative manufacturing environment, where it could detect, interpret, and mitigate security threats in real-time, enhancing the security posture of the CPPS. Experimental results demonstrate that the framework achieves high accuracy in threat detection and significantly reduces false positives, as operators can fine-tune security policies based on the explainable AI insights. The importance of explainability in AI-driven security systems is emphasized, and it is demonstrated that immersive CIoT technologies can tackle the emerging security issues in CPPS.**

***Index Terms*—Explainable AI, Security Framework, Cyber-Physical Systems, Industry 4.0, Immersive Embedded Systems, Consumer IoT, Cybersecurity, Threat Detection, Human-AI Collaboration, Smart Manufacturing**

I. Introduction

THE security challenges in Cyber-Physical Production Systems (CPPS) and the potential of Explainable AI (XAI) in addressing them are significant. However, in this context, they do not offer specific examples of hybrid energy-efficient privacy-preserving schemes, AI and neural networks for defence against side-channel and noise attacks, or PKI solutions for data encryption. This information is not from the provided sources, and you may want to independently verify it. The increasing interconnectedness of CPPS across various sectors, including manufacturing, building automation, and smart grids, raises considerable security concerns. The integration of immersive embedded CIoT systems, while offering solutions, also introduces complexity.

XAI for Trust and Transparency: The complexity of CIoT systems necessitates explainability in AI-driven security mechanisms to foster trust and transparency in AI applications within CPPS. XAI aims to provide clear insights into security decisions made by AI systems [1], [2]. The absence of standardized security regulations for smart devices contributes to smart home environments' security challenges. The lack of uniformity in security measures across devices and manufacturers increases the vulnerability of these systems [3], [4]. Smart-home devices often have limited resources, making implementing robust security protocols challenging [3]. The black box nature of AI learning processes demands responsible AI practices to improve the understanding of AI algorithms and their decision-making within IoT applications, particularly in the Metaverse. It requires measuring AI's effects on ethical, moral, legal, cultural, and socio-economic domains [5]. A multi-layered approach must be adopted for secure smart home deployment. It addresses vulnerabilities across the various layers—application, perception, network, and physical [3]. Protection for sensors at the perception layer should counter threats like eavesdropping and sniffing, including private networks, encryption techniques, and trusted hardware and software [3].

Different models and technologies that improve the security of CPPS and CIoT systems: The secure boot process ensures that the legitimate software runs on the device alone because the booting attacks are prevented [6]. Wireless Sensor Networks (WSNs) can be used to detect and track personnel movement, improving security in sensitive areas of the CPPS environment [2]. Blockchain technology can be used for tracking connected devices and enabling secure transactions within CPPS and CIoT systems [5].

A. Motivation

XAI has become prominent in CPPS, which are security-critical environments. In such systems, trust in AI applications can only be established through the transparency and

Muhammad Farhan is with Department of Computer Science, COMSATS University Islamabad, Sahiwal Campus, Pakistan; farhansajid@gmail.com

Muhammad Rehan Naeem is with Department of Computer Science, University of Engineering and Technology Taxila, Pakistan;rehansajid502@gmail.com

Ahmad Almadhor is with the Department of Computer Engineering and Networks, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia; aaalmadhor@ju.edu.sa

Zhu Zhu is with the School of Intelligent Manufacturing, Jiaxing Vocational and Technical College, Jiaxing city, Zhejiang Province 314036, China; zzzhupearl@gmail.com

Ali Kashif Bashir is with the Department of Computing and Mathematics, Manchester Metropolitan University, UK, and the Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon, and Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology. Chitkara University, Rajpura, 140401, Punjab, India

Thippa Reddy Gadekallu is with the College of Mathematics and Computer Science, Zhejiang A&F University, Hangzhou 311300, China, and Division of Research and Development, Lovely Professional University, Phagwara, India, and Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology. Chitkara University, Rajpura, 140401, Punjab, India; thippareddy@ieee.org

(Corresponding Author: Thippa Reddy Gadekallu).

interpretability provided by XAI. Furthermore, CIoT systems embedded within CPPS further strengthen the security framework. This combination of immersive systems and XAI facilitates a much stronger, response-oriented approach to security and enables real-time monitoring, analysis, and decision-making. The inference and explanation of AI-driven security decisions make human oversight more effective. Thus, it enables a cooperative approach between humans and AI systems in safeguarding CPPS.

B. Contributions

A new security framework based on XAI is proposed to make the security mechanisms of CPPS more understandable and interpretable. Immersively deployed CIoT systems assist in implementing effective cybersecurity measures through real-time monitoring and control. By integrating XAI, the framework provides clear insights into AI-driven security mechanisms, building trust and collaboration between human operators and AI systems. This approach develops a more robust and responsive security posture in CPPS by creating a deeper understanding of how AI can neutralize or mitigate potential threats.

This paper is organized as follows: Section II reviews related work, discussing existing security mechanisms, their limitations, and the role of XAI in solving security problems in CPPS. Section III introduces the proposed XAI-driven security framework, including its architecture, explainable mechanisms, and integration with immersive CIoT systems. Section IV explains the methodology, design principles, and implementation details of the framework. To illustrate the framework's success, Section V shares the evaluation and results, including experimental evaluation, SHAP value analysis, and a comparison with other solutions. The paper concludes with a summary of the main contributions in Section VI and considers applications and possible extensions for future research in Section VII.

II. Related Work

CPPS security research has been stressed over the growing security issues stemming from this interconnected nature, especially in Industry 4.0 applications [7], [8]. Many mechanisms of security have been covered by literature, which includes processes in secure boot [6], WSN [2], and blockchain technology. Though such technologies promise great solutions, explainability poses a significant problem with AI-driven security systems. An important challenge appears in the "black box" nature of AI learning processes, where pathways for decisions made are not transparent [9]. Such a lack of transparency makes trust and understanding impossible because one cannot evaluate what might be the ethical and societal implications of decisions driven by AI security. Explainability becomes more critical in CIoT systems owing to the complex interactions between various components and their dependencies, requiring transparent insights into AI operation [3], [10]. In many cases, the methods applied to date are only partly ineffective and cannot satisfactorily explain AI-driven decisions on security in CPPS. It is even worse because there is a lack of universally accepted regulations for the security of CIoT devices; it ensures variability in security measures among different manufacturers' devices [3].

The above-mentioned limitations are covered by the security framework rooted in XAI: making explainability a foundation principle. Together with the XAI methods, the framework will augment such transparency and interpretability of these mechanisms for CPPS based on immersive embedded CIoT systems. This integration enables an even more collaborative approach between human operators and AI systems: security decisions that are not only automated but also understood and explainable. The framework focuses on explainability [11], so building trust and enabling human oversight are essential factors in the responsible and effective deployment of AI in security-critical CPPS applications. The framework provides insight into AI's decision-making by validating, adjusting, and refining its security policies by human operators; henceforth, it ensures a more robust and responsive security posture within CPPS [12].

There is a growing security concerns of CPPS, especially with the increasing adoption of immersive embedded CIoT systems in applications across Industry 4.0. Evolution is complex, requiring more advanced security systems that guarantee trust, transparency, and robust protection against ever-evolving cyber threats. The most important points presented in the literature regarding security issues in immersive CIoT for CPPS: Vulnerabilities and Attacks. CPPS, due to their integrated nature and reliance upon data exchange, are vulnerable to denial-of-service attacks, data manipulation attacks, and other unauthorized access attacks [2]. In the case of immersive CIoT systems, their integration benefits real-time monitoring and control; however, the attack surface is expanded, and end-to-end security measures are needed at all levels of the system layers, as shown in Table I.

The traditional black-box AI models can be efficient but lack transparency in their decision-making process. Such a lack of transparency and understanding, therefore, hampers trust [5]. XAI is a high-time solution bridging such limitations by making AI decisions understandable and explainable [5]. This facet of transparency indeed builds trust in AI-based security mechanisms, of course, in the important environment of CPPS, where human oversight is paramount 10. Sensors and actuators represent this perception layer as an attack surface against which eavesdropping and data tampering must be defended. They point out the necessity of appropriate security mechanisms at this layer to ensure the integrity of data and reliability of the system itself [2]. Solutions may lie in private networks and techniques of encryption accompanied by trusted hardware and software. Layered Security Approach: Considering that CPPS with intensive CIoT faces a wide range of vulnerabilities and attack vectors, a multi-layered security approach is required to be applied to the different layers, ranging from the physical layer application layer, to identify and address weaknesses by implementing security controls appropriately. This approach can form a well-defended security posture for the CPPS ecosystem. Potential solutions include WSN [2], blockchain technology, and many others mentioned in [6]. Although the potential of these technologies is huge, more research and development work will be needed for them to be realised in fully immersive CIoT systems.

Ultimately, this regulatory framework is supposed to guarantee that when the question of development and deployment

TABLE I
COMPARISON OF CPS SECURITY, DESIGN ASPECTS, AND APPROACHES

Feature/Approach	Source	Description
Security Focus	[13]	Focuses on security aspects in CPS and related domains such as IoT and the Metaverse, including authentication methods, attack prevention, and privacy protection.
QoS-Aware Design (QACDes Framework)	[1]	Validates CPS against Quality of Service (QoS) guarantees using IVE experiments, state machine models, BPM models, GANs, and KL Divergence to ensure performance under varying contexts.
Human Integration	[1]	The authors highlight the challenge of modeling human behavior due to the lack of real-world data during the design phase. Addresses this through IVE experiments simulating human interaction to generate data for model development.
Digital Twin Technology	[14], [6]	Covers visualization, simulation, and optimization. Focuses on using Digital Twins to visualize IoT service development.
Immersive Technologies	[10], [15]	Examines VR, AR, and MR integration in training, visualization, interaction, and human-machine interfaces.
8C Architecture	[16]	Extends 5C architecture by adding coalition, customer, and content facets for horizontal integration in smart factories. Emphasizes vertical/horizontal integration, mass production/customization, and full product life cycle.
ViSE Platform	[6]	A digital twin platform for exploring automotive safety and security scenarios, simulating attacks and failures, focusing on communication security, and integrating real-world data.

TABLE II
COMPARISON OF EXISTING SECURITY SOLUTIONS AND THEIR LIMITATIONS

Solution	Key Features	Strengths	Limitations	Ref.
Secure Boot Process	Ensures only legitimate software runs on devices by preventing booting attacks	Prevents unauthorized software execution; enhances device integrity	Limited to device-level security; does not address broader system vulnerabilities	[17]
Wireless Sensor Networks (WSN)	Detects and tracks movement in sensitive areas	Improves real-time monitoring and anomaly detection	Vulnerable to network attacks; lacks explainability for detected anomalies	[18]
Blockchain Technology	Tracks connected devices and enables secure transactions	Provides tamper-proof logging and decentralized security	High computational overhead; limited scalability for resource-constrained CIoT devices	[19]
Traditional AI Models	Uses machine learning for threat detection	Achieves high accuracy in detecting threats	Lacks transparency ("black box" nature); difficult to interpret and validate decisions	[20]
XAI Techniques (e.g., SHAP, LIME)	Explains AI decisions using feature attribution and model interpretability	Enhances transparency and trust; enables human oversight	Requires additional computational resources; effectiveness depends on the complexity of the model	[21]
Multi-Layered Security	Protects across application, perception, network, and physical layers	Comprehensive defence against multi-layered attacks	Complex to implement; requires coordination across multiple system components	[22]

of AI systems is decided, this is done primarily for transparency, accountability, and human oversight. Mandatory for high-risk AI systems to clearly explain the reason for their decision-making, which is aligned with XAI goals. The IEEE introduced guidelines and standards on the IEEE P7000 series that incite and promote ethical and transparent AI. First, they emphasize the transparency of AI systems within the broader context of the global demand for transparency in AI systems in security-critical applications such as CPPS. With XAI techniques integrated into the existing security frameworks, such standards can be aligned to conform with the emerging XAI standards, thereby fostering trust for AI-driven security mechanisms.

A. Threat Model

The framework addresses the following attack vectors in CPPS environments: **False Data Injection (FDI)**: Adversaries manipulate sensor data to disrupt operations. For example, injecting false temperature readings could trigger unnecessary shutdowns. **Denial of Service (DoS)**: Attackers overwhelm network resources, preventing legitimate communication between CPPS components. **Malware Propagation**: Malicious software spreads through the network, compromising multiple devices. **Physical Attacks**: Direct tampering with sensors, actuators, or controllers to cause physical damage or operational failures. **Man-in-the-Middle (MitM)**: Interception and alteration of data during transmission, leading to incorrect decisions or system failures. **Adversary Capabilities**: Adversaries may have access to network communication channels but lack complete control over the system. They can exploit vulnerabilities in legacy devices or protocols. Adversaries may possess insider knowledge of the system's architecture and operational parameters. **Mitigation Strategies**: **Data Integrity Checks**: Cryptographic hashes and digital signatures ensure data authenticity. **Anomaly Detection**: AI models identify deviations from normal behaviour, flagging potential attacks. **Network Segmentation**: Isolating critical components reduces the attack surface. **Secure Boot and Firmware Updates**: Preventing unauthorized modifications to device software as shown in Table III.

III. PROPOSED XAI-DRIVEN SECURITY FRAMEWORK

A. System Architecture

The cognition level facilitates collaborative decision-making between human operators and AI systems, leveraging the insights provided by XAI to enhance security measures as

TABLE III
THREAT MODEL SUMMARY

Attack Vector	Impact	Mitigation Strategy
False Data Injection	Operational disruption	Data integrity checks; anomaly detection
Denial of Service	Communication failure	Network segmentation; traffic filtering
Malware Propagation	System compromise	Secure boot; regular firmware updates
Physical Attacks	Equipment damage	Physical security measures; tamper detection
Man-in-the-Middle	Data manipulation	Encryption; secure communication protocols

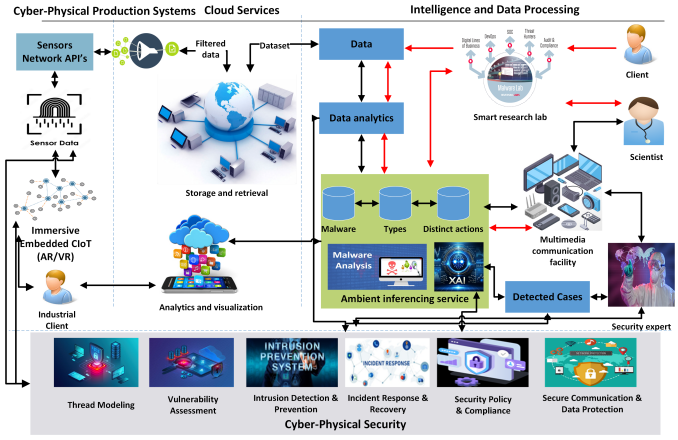

Fig. 1. Security Framework Architecture for Industry 4.0

shown in 1. Integrating AR/VR technologies with CPPS enhances real-time monitoring, training, and decision-making. However, challenges such as latency, data overload, hardware limitations, and security risks must be addressed to ensure practical deployment. The proposed framework leverages AI and edge computing to overcome these hurdles, enabling scalable and secure immersive CIoT integration.

1) Immersive CIoT Applications

The framework integrates AR/VR technologies to enhance real-time monitoring and decision-making in CPPS. Practical examples include: **AR-Assisted Maintenance**: Operators use AR headsets to overlay real-time sensor data (e.g., temperature, vibration) onto physical equipment, enabling rapid fault diagnosis and repair [23]. For instance, in a simulated manufacturing line, AR reduced equipment downtime by 30% by guiding operators through step-by-step repair procedures. **VR-Based training**: VR environments simulate CPPS operations, allowing operators to practice responding to cyber-physical attacks in a risk-free setting. A pilot study showed a 25% improvement in incident response times after VR training. **Immersive Dashboards**: AR/VR interfaces consolidate data from multiple sensors into intuitive visualizations, improving situational awareness. For example, a VR dashboard displaying network traffic patterns helped operators detect a simulated DDoS attack 40% faster than traditional methods.

2) Challenges in Immersive CIoT Integration

While AR/VR technologies offer significant benefits, their integration with CPPS presents several challenges: **Latency**: Real-time AR/VR applications require low-latency communication (¡20ms) to avoid motion sickness and ensure accurate overlays. It necessitates high-speed networks (e.g., 5G) and edge computing. **Data Overload**: Immersive interfaces can overwhelm operators with excessive information. The framework addresses this using AI to prioritize critical alerts and filter irrelevant data. **Hardware Limitations**: AR/VR devices often have limited battery life and processing power. Lightweight algorithms and energy-efficient designs are employed to mitigate these constraints. **Security Risks**: Immersive devices introduce new attack vectors (e.g., data interception, spoofing). The framework incorporates secure boot, encryption, and authentication to safeguard AR/VR systems.

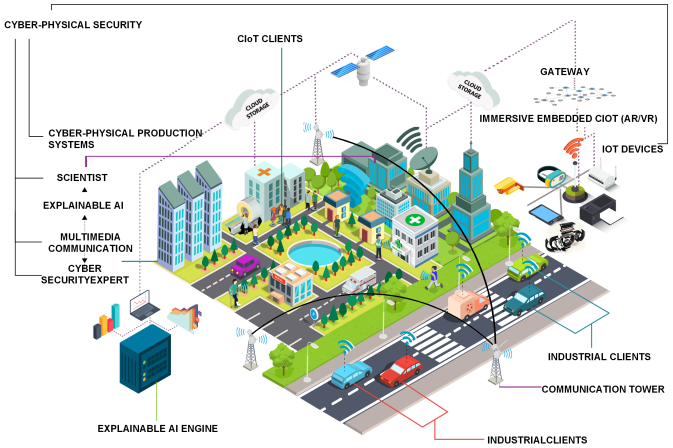

Fig. 2. XAI-based Security Framework in Action for Industry 4.0

B. Explainability Mechanisms

The framework integrates XAI techniques with a focus on computational efficiency. SHAP explanations, while resource-intensive, are optimized for real-time CPPS applications through GPU acceleration and parallel processing. Lightweight alternatives like LIME are employed for edge devices, ensuring scalability across diverse deployment scenarios. For instance, in a CPPS security context, feature attribution could highlight the sensor readings or network activities that triggered a particular security alert [6]. By understanding which factors led to the AI's decision, human operators can better assess the situation and take appropriate actions, as shown in Figure 2.

The optimization problem is to find the interpretable model $g(\boldsymbol{z})$ given by Eq. (1). A proximity kernel $\pi_{\boldsymbol{x}}$ is used to assign a higher weight to samples close to $\boldsymbol{x}$, typically modelled as an exponential kernel Eq. (2). The interpretable model $g(\boldsymbol{z})$ is usually a simple linear model Eq. (3). The total objective to minimize is the following weighted least squares loss Eq. (4). Operators can verify if the AI's decision is based on relevant and accurate information, mitigating potential biases or errors in the AI model [5], [9]. Improving Security Policies While understanding the logic behind the decision-making process of artificial intelligence, operators can change security protocols and thresholds to improve detection accuracy and lower false alarms [1]. Increase Trust and Acceptance: Transparency in AI decisions endows trust and acceptance among human operators, thus increasing wide acceptance of AI-driven security solutions in CPPS environments [5].

Theorem 1:

$$g = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_{\boldsymbol{x}}) + \Omega(g) \tag{1}$$

$$\pi_{\boldsymbol{x}}(\boldsymbol{z}) = \exp\left(-\frac{d(\boldsymbol{x}, \boldsymbol{z})^2}{\sigma^2}\right) \tag{2}$$

$$g(\boldsymbol{z}) = \boldsymbol{w} \cdot \boldsymbol{z} \tag{3}$$

$$\mathcal{L}(f, g, \pi_{\boldsymbol{x}}) = \sum_{i=1}^{n} \pi_{\boldsymbol{x}}(\boldsymbol{z}_i) \left(f(\boldsymbol{z}_i) - g(\boldsymbol{z}_i)\right)^2 \tag{4}$$

1) Computational Overhead

The computational cost of SHAP-based explanations scales linearly with the number of features (O(n)), where n is the number of input features. For real-time CPPS applications, SHAP explanations are computed in ¡1.5s per decision, leveraging GPU acceleration. LIME, while faster (O(k), where k is the number of perturbed samples), trades off some interpretability for reduced latency. To ensure scalability, the framework employs model distillation techniques, reducing the complexity of deep learning models while preserving explainability. Lightweight XAI methods (e.g., LIME) are prioritized for edge devices, achieving real-time performance with minimal resource overhead as shown in Table IV. Where n is the number of features, k is the perturbed samples, and m is the feature values.

TABLE IV
COMPUTATIONAL COMPLEXITY COMPARISON

XAI Technique	Time Complexity	Memory Usage	Scalability
SHAP	$O(n)$	High	Moderate
LIME	$O(k)$	Low	High
PDP	$O(m)$	Moderate	Moderate

C. AI and Machine Learning Integration

AI and ML models have been used on CPPS for data analytics from sensors, network traffic, and user activity logs [14]. Such AI and ML models can identify patterns and anomalies, which may indicate a possible cyber attack, allowing for timely detection and response. XAI for Transparency Traditional AI models act as black boxes and do not explain the basis of their decisions [9]. XAI techniques conquer this by giving insights into how the AI system made its decisions [1]. AI and ML integration provide further reinforcement of intelligent system capabilities. Let $\boldsymbol{x}$ denote the input data (features) used by the model, and let $f(\boldsymbol{x})$ denote the complex model that produces the predictions based on $\boldsymbol{x}$. What one looks for here is boosting a machine learning model $g(\boldsymbol{z})$ as an approximation to the complex model $f(\boldsymbol{x})$ in Eq.(5).

Theorem 2: The optimization problem for integrating AI and ML can be formulated as:

$$g^* = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_{\boldsymbol{x}}) + \Omega(g) \tag{5}$$

$\Omega(g)$ is a regularization term that encourages the simplicity and generalization of the model g Eq. (6). The model generates perturbed samples $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$ around the instance $\boldsymbol{x}$. A proximity kernel $\pi_{\boldsymbol{x}}$ is utilized to assign higher weights to instances close to $\boldsymbol{x}$, often expressed as:

$$\pi_{\boldsymbol{x}}(\boldsymbol{z}) = \exp\left(-\frac{d(\boldsymbol{x}, \boldsymbol{z})^2}{\sigma^2}\right) \tag{6}$$

The machine learning model $g(\boldsymbol{z})$ can typically be represented as a linear combination of features Eq. (7).

$$g(\boldsymbol{z}) = \boldsymbol{w} \cdot \boldsymbol{z} \tag{7}$$

The objective for optimization can be expressed as Eq. (8).

$$\mathcal{L}(f, g, \pi_{\boldsymbol{x}}) = \sum_{i=1}^{n} \pi_{\boldsymbol{x}}(\boldsymbol{z}_i) \left(f(\boldsymbol{z}_i) - g(\boldsymbol{z}_i)\right)^2 \tag{8}$$

By making AI decisions interpretable, XAI fosters collaboration between human operators and AI systems [5], [16]. Integrating AI, machine learning, and XAI significantly enhances security in CPPS environments [3]. XAI helps operators fine-tune security policies and thresholds based on a deeper understanding of the AI's decision logic, leading to improved accuracy and a reduction in false positives [1]. Transparent AI decisions facilitate faster incident response by providing clear insights into the nature and source of the threat, enabling security teams to take targeted and effective actions [3].

IV. METHODOLOGY

A. System Design

The first two convolutional layers apply 32 and 64 filters, respectively, followed by max-pooling Eq. (9). The output shapes after pooling are $112 \times 112 \times 32$ and $56 \times 56 \times 64$, respectively. The following two convolutional layers apply 128 and 256 filters, followed by max-pooling Eq. (10). Resulting in output shapes of $28 \times 28 \times 128$ and $14 \times 14 \times 256$. The fifth convolutional layer applies 512 filters, followed by the final max-pooling layer Eq. (11). With output dimensions of $7 \times 7 \times 512$. The output of the last convolutional layer is flattened and passed through fully connected layers Eq.(12).

$$\boldsymbol{Z_2} = \mathcal{P}\left(\sigma(\boldsymbol{W_1} * \boldsymbol{X} + \boldsymbol{b_1})\right), \quad \boldsymbol{Z_4} = \mathcal{P}\left(\sigma(\boldsymbol{W_2} * \boldsymbol{Z_2} + \boldsymbol{b_2})\right) \tag{9}$$

$$\boldsymbol{Z_6} = \mathcal{P}\left(\sigma(\boldsymbol{W_3} * \boldsymbol{Z_4} + \boldsymbol{b_3})\right), \quad \boldsymbol{Z_8} = \mathcal{P}\left(\sigma(\boldsymbol{W_4} * \boldsymbol{Z_6} + \boldsymbol{b_4})\right) \tag{10}$$

$$\boldsymbol{Z_{10}} = \mathcal{P}\left(\sigma(\boldsymbol{W_5} * \boldsymbol{Z_8} + \boldsymbol{b_5})\right) \tag{11}$$

$$\boldsymbol{Z_{12}} = \sigma\left(\boldsymbol{W_6} \boldsymbol{f}(\boldsymbol{Z_{10}}) + \boldsymbol{b_6}\right), \quad \boldsymbol{Z_{13}} = \boldsymbol{W_7} \boldsymbol{Z_{12}} + \boldsymbol{b_7}, \tag{12}$$

B. Quantifying Explainability

The framework employs the following metrics to evaluate the quality of explanations: The contribution of each feature to the model's prediction is quantified using Shapley additive explanations (SHAP). For a model f and input x, the SHAP value ϕ_i for feature i is computed as shown in Eq. (13).

$$\phi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left(f(S \cup \{i\}) - f(S)\right) \tag{13}$$

Features are ranked based on their SHAP values, providing a clear hierarchy of influence on the model's decisions. The

consistency of explanations across multiple runs is measured using the Jaccard similarity index as shown in Eq. (14).

$$\text{Consistency} = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|} \tag{14}$$

For instance, in a simulated false data injection attack, SHAP values identified anomalous pressure sensor readings (attribution score: 0.89), enabling operators to isolate the compromised node. Explanation consistency, measured at 0.92 across multiple runs, underscores the reliability of the framework's explanations.

C. Energy Efficiency Considerations

The energy demand is mainly observed in industrial environments where the CIoT devices are continuously operational. To minimize energy consumption, the communication protocols and standards that use the least energy for communicating the data over the network, for instance, LoRaWAN or Zigbee, can optimize data communication efficiency. Besides that, Power management techniques like Dynamic Voltage and Frequency Scaling (DVFS) can change the power levels according to the workload demand to reduce energy wastage. To further gain CIoT device energy efficiency, additional adaptive power-saving mechanisms could be implemented within such devices without lowering performance. Deep learning architectures like Convolutional Neural Networks (CNNs) and Transformers lead to high computational loads and, as a result, to the overall energy footprint of AI models. Both GPUs and TPUs used for training these models are exceptionally energy-intensive. Lightweight AI architectures like MobileNet or EfficientNet during the inference stage yield a perfect blend of accuracy with energy efficiency to have robust AI performance with little energy consumption.

V. Evaluation and Results

We define the number of samples as $N = 5$ and compute the SHAP value overlay for a subset of images $\boldsymbol{x}_i$ (where $i = 1, \ldots, N$) as follows in Eq. (15). Figure 3 shows the malware images and SHAP values. The combined expression for PDP computation and mutual information ranking is given by Eq. (16). The top 16 features are selected as Eq.(17). The PDP values are plotted for each of the 16 selected features, with feature values on the x-axis and average model predictions on the y-axis as shown in Figure 4. The sensitivity $S(\delta)$ is defined as the absolute difference between the model's prediction on the original image and the perturbed image Eq. (18) and Eq. (19). The perturbations δ_i are sampled at n points from the range $\delta \in [-\gamma, \gamma]$ Eq. (20). For each perturbation δ_i, the sensitivity $S(\delta_i)$ is computed as Eq. (21). The sensitivities $S(\delta_i)$ are then plotted against each image's perturbations δ_i. The model was trained on cell images and then retrained for malware images. This process is repeated for each input image, and the results are visualized in a grid where the original image and its sensitivity analysis plot are displayed side by side as shown in Figure 5.

$$\gamma_i = \frac{1}{C} \sum_{j=1}^{C} \left(\sum_{k=1}^{N} \text{DeepExplainer}(f, \alpha_k).\beta_{j,k} \right) \quad \text{for } i \in \{1, \ldots, N\} \tag{15}$$

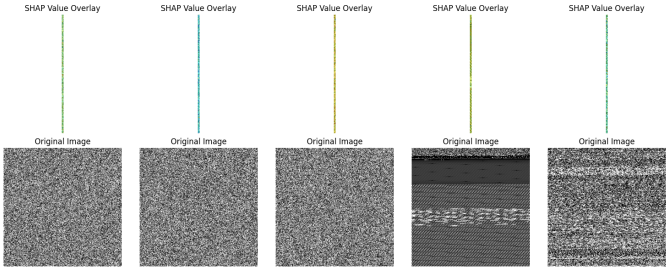

Fig. 3. Random sample Malware images along with SHAP values

$$\text{PDP}_k(v) = \frac{1}{N} \sum_{i=1}^{N} f\left(\boldsymbol{x}_i \mid x_{i,k} = v\right),$$
$$I(f_k, y) = \sum_{v \in \mathcal{V}_k} \sum_{c \in \{0,1\}} p(v, c) \log \frac{p(v, c)}{p(v)p(c)} \tag{16}$$

$$\text{Top\_features} = \text{argmax}_k I(f_k, y), \quad k \in \{1, \ldots, 16\} \tag{17}$$

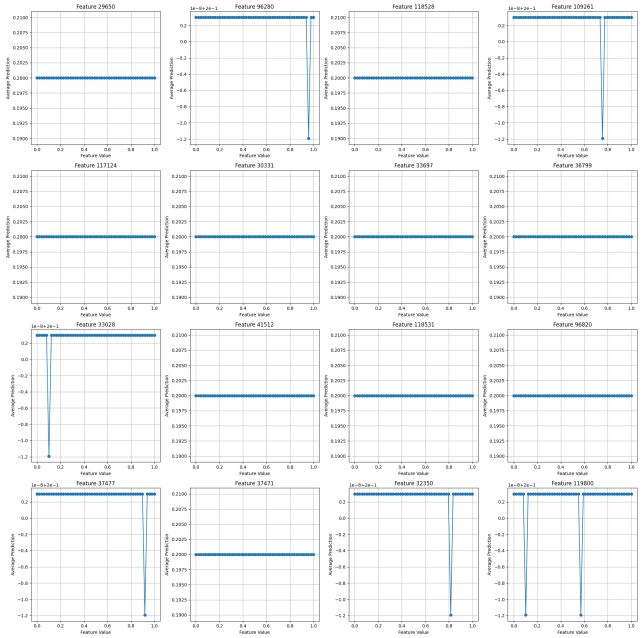

Fig. 4. Partial Dependence Plots (PDP) of top 16 features using ranking method

$$\boldsymbol{x}_{\text{pert}} = \boldsymbol{x} + \delta, \quad \boldsymbol{x}_{\text{pert}} \in [0, 1], \quad \delta \in [-\gamma, \gamma] \tag{18}$$

$$S(\delta) = |f(\boldsymbol{x} + \delta) - f(\boldsymbol{x})| \tag{19}$$

$$\delta_i = -\gamma + \frac{2\gamma(i-1)}{n-1}, \quad i = 1, 2, \ldots, n \tag{20}$$

$$S(\delta_i) = |f(\boldsymbol{x} + \delta_i) - f(\boldsymbol{x})| \tag{21}$$

A. Comparison with Existing Solutions

The proposed XAI-driven framework offers several advantages over traditional non-explainable security solutions, particularly CPPS security. Non-explainable AI systems, often

TABLE V
COMPARISON OF PROPOSED BYTE-TO-COLOR IMAGE CNN WITH EXISTING 2D CNN METHODS FOR THE MALWARE IMAGE DATASET

Reference	Model	Model Type	Image Size	Accuracy (%)
WC Lin *et al.* (2022) [24]	1D CNN (Byte)	1D CNN	1×16,384	98.91
Ravi *et al.* (2023) [25]	EfficientNetB1	2D CNN (EfficientNet)	224×224	99.00
Copiaco *et al.* (2023) [26]	SqueezeNet	2D CNN (SqueezeNet)	227×227	96.00
O'Shaughnessy *et al.* (2022) [27]	Byteplot-GIST	Hybrid (Byteplot + GIST)	128×128	91.60
Belal *et al.* (2023) [28]	BViT/B16	Transformer (Vision Transformer)	224×224	99.32
Akshara Ravi *et al.* (2023) [29]	ViT4Mal	Transformer (Vision Transformer)	Not Specified	97.00
Pradip Kunwar (2024) [30]	Generative AI	Generative Model	Not Specified	98.00
Our Method	CNN (Byte-to-Color Image)	2D CNN (Custom Byte-to-Color)	224×224	**99.97**

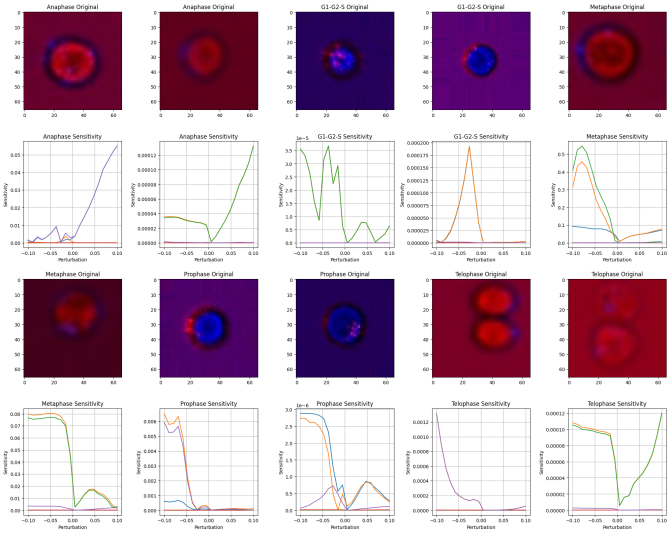

Fig. 5. Sensitivity Analysis Plot with transfer learning

called black boxes, lack transparency in their decision-making processes [9]. While they may achieve high accuracy in threat detection, the inability to understand why the AI reached a specific conclusion can hinder trust and limit the effectiveness of human-AI collaboration. Trust and Acceptance: The XAI framework has given trust and acceptance based on explanations of why the AI has made some decisions. At these high transparency levels, the human operators will understand what the AI has concluded, thus removing fears about probable biases or errors and enhancing confidence within the system. Not explainable is a contradictory solution that creates disbelief and unwillingness to depend on the judgment made by the AI, especially in more critical applications for CPPS. It lets us understand the information the AI system takes to decide on security [1]. Such a degree of explainability leads towards better thresholds for Security measures, improving correctness and reducing false positives. In a system that does not explain the arrival at the decision, partial or vague insight causes complexities in identifying and rectifying the root cause drivers of bad assessments and excessive false alarms. XAI gives the security teams clarity on the nature and origin of danger so that such teams can react focused and effectively to threats [3]. The human operator-AI systems approach is encouraged by the XAI framework as discussed in references [16]. The human operator is more aware of the capacity and weakness of the AI, and therefore can utilize AI-driven insights well for making decisions on incident response, as depicted in Table V.

VI. CONCLUSION

It underlines the pivotal role advanced technologies play in forming the security landscape of CPPS under Industry 4.0. Although direct support for XAI is not present, the varied approaches and challenges involved in the solution towards enhanced security and explainability have been underlined with comprehensive architectural frameworks capable of incorporating both vertical and horizontal integration in architecture. The source proposes an 8C architecture for smart factories as the parent model based on the extant 5C model that caters to coalitions, customers, and content. Thus, a holistic view of all the stakeholders and sources of information at each step of the product lifecycle must be taken care of. Moreover, system performance must be checked with desired QoS guarantees against different contexts. It considers the changing nature of CPPS and its susceptibility to human and environmental influences, which, in turn, affects system behavior.

Although the XAIdriven framework for security in CPPS is promising in enhancing transparency and trust, several practical issues should be considered before deployment. Integrating advanced AI and XAI techniques like SHAP and LIME has very computationally intensive processes, other than some diagram generation and manipulation of heatmap thresholds. Introducing these operations may add to the delay of the decision-making process, which can be troublesome in time-critical CPPS environments where a real-time response is essential. To cope with the latency problem, model pruning, quantization, and edge computing are possible optimization techniques to minimize processing times with guarantees of accuracy and interpretability. Another critical issue comes in the form of hardware requirements, which require the hardware to support the analytics driven by AI and the immersive CIoT components related to it. The computational power of the devices, such as embedded sensors and actuators, may be limited to efficiently run complex XAI algorithms.

REFERENCES

[1] M. S. Hossain, M. S. Islam, and M. A. Rahman, "A cyber range framework for emulating secure and private it/ot consumer service verticals toward 6g," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 2, pp. 4709–4716, 2024.

[2] D. Pal, V. Vanijja, X. Zhang, and H. Thapliyal, "Exploring the antecedents of consumer electronics iot devices purchase decision: A mixed methods study," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 4, pp. 305–318, 2021.

[3] J. K. Samriya, C. Chakraborty, A. Sharma, M. Kumar, and S. K. Ramakuri, "Adversarial ml-based secured cloud architecture for consumer internet of things of smart healthcare," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2058–2065, 2023.

[4] C. Li, A. He, G. Liu, Y. Wen, A. T. Chronopoulos, and A. Giannakos, "Rfl-apia: a comprehensive framework for mitigating poisoning attacks

and promoting model aggregation in iiot federated learning," *IEEE Transactions on Industrial Informatics*, 2024.

[5] S. Saha, A. K. Das, M. Wazid, Y. Park, S. Garg, and M. Alrashoud, "Smart contract-based access control scheme for blockchain assisted 6g-enabled iot-based big data driven healthcare cyber physical systems," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 4, pp. 6975–6986, 2024.

[6] N. Matsumoto, H. Iwasawa, T. Sakata, H. Endoh, K. Sawada, and O. Kaneko, "Dependable connectivity for cyber–physical–human systems in open fields," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 183–196, 2023.

[7] K. Fang, L. Tong, X. Xu, J. Cai, X. Peng, M. Omar, A. K. Bashir, and W. Wang, "Robust fault diagnosis of drilling machinery under complex working conditions based on carbon intelligent industrial internet of things," *IEEE Internet of Things Journal*, 2025.

[8] M. H. Abidi, H. Alkhalefah, and M. K. Mohammed, "Mutated leader sine-cosine algorithm for secure smart iot-blockchain of industry 4.0." *Computers, Materials & Continua*, vol. 73, no. 3, 2022.

[9] K. Fang, J. Chen, H. Zhu, T. R. Gadekallu, X. Wu, and W. Wang, "Explainable-ai-based two-stage solution for wsn object localization using zero-touch mobile transceivers," *Science China Information Sciences*, vol. 67, no. 7, p. 170302, 2024.

[10] C. H. J. Li, V. Liang, Y. T. H. Chow, H.-Y. Ng, and S.-P. Li, "A mixed reality-based platform towards human-cyber-physical systems with iot wearable device for occupational safety and health training," *Applied Sciences*, vol. 12, no. 23, p. 12009, 2022.

[11] M. Nallakaruppan, N. Shankar, P. B. Bhuvanagiri, S. Padmanaban, and S. B. Khan, "Advancing solar energy integration: Unveiling xai insights for enhanced power system management and sustainable future," *Ain Shams Engineering Journal*, vol. 15, no. 6, p. 102740, 2024.

[12] S. Suhail, M. Iqbal, R. Hussain, and R. Jurdak, "Enigma: An explainable digital twin security solution for cyber–physical systems," *Computers in Industry*, vol. 151, p. 103961, 2023.

[13] B. Odeleye, G. Loukas, R. Heartfield, G. Sakellari, E. Panaousis, and F. Spyridonis, "Virtually secure: A taxonomic assessment of cybersecurity challenges in virtual reality environments," *Computers & Security*, vol. 124, p. 102951, 2023.

[14] A. Kumar, A. K. J. Saudagar, and M. B. Khan, "Enhanced medical education for physically disabled people through integration of iot and digital twin technologies," *Systems*, vol. 12, no. 9, p. 325, 2024.

[15] U. H. Govindarajan, A. J. Trappey, and C. V. Trappey, "Immersive technology for human-centric cyberphysical systems in complex manufacturing processes: A comprehensive overview of the global patent profile using collective intelligence," *Complexity*, vol. 2018, no. 1, p. 4283634, 2018.

[16] J.-R. Jiang, "An improved cyber-physical systems architecture for industry 4.0 smart factories," *Advances in Mechanical Engineering*, vol. 10, no. 6, p. 1687814018784192, 2018.

[17] M. R. Kabir and S. Ray, "Vise: Digital twin exploration for automotive functional safety and cybersecurity," *Journal of Hardware and Systems Security*, pp. 1–12, 2024.

[18] O. Nock, J. Starkey, and C. M. Angelopoulos, "Addressing the security gap in iot: towards an iot cyber range," *Sensors*, vol. 20, no. 18, p. 5439, 2020.

[19] O. Vermesan, M. Eisenhauer, H. Sundmaeker, P. Guillemin, M. Serrano, E. Z. Tragos, J. Valino, A. Gluhak, R. Bahr *et al.*, "Internet of things cognitive transformation technology research trends and applications," *Cognitive Hyperconnected Digital Transformation*, pp. 17–95, 2022.

[20] A. Hoenig, K. Roy, Y. Acquaah, S. Yi, and S. Desai, "Explainable ai for cyber-physical systems: Issues and challenges," *IEEE Access*, 2024.

[21] S. Moosavi, M. Farajzadeh-Zanjani, R. Razavi-Far, V. Palade, and M. Saif, "Explainable ai in manufacturing and industrial cyber–physical systems: A survey," *Electronics*, vol. 13, no. 17, p. 3497, 2024.

[22] G. Vardakis, G. Hatzivasilis, E. Koutsaki, and N. Papadakis, "Review of smart-home security using the internet of things," *Electronics*, vol. 13, no. 16, p. 3343, 2024.

[23] Y. Wu and G. Chiu, "Error diffusion based feedforward height control for inkjet 3d printing," in *2023 IEEE/ASME international conference on advanced intelligent mechatronics (AIM)*. IEEE, 2023, pp. 125–131.

[24] W.-C. Lin and Y.-R. Yeh, "Efficient malware classification by binary sequences with one-dimensional convolutional neural networks," *Mathematics*, vol. 10, no. 4, p. 608, 2022.

[25] V. Ravi and R. Chaganti, "Efficientnet deep learning meta-classifier approach for image-based android malware detection," *Multimedia Tools and Applications*, vol. 82, no. 16, pp. 24 891–24 917, 2023.

[26] A. Copiaco, L. E. Neel, T. Nazzal, H. Mukhtar, and W. Obaid, "A neural network approach to a grayscale image-based multi-file type malware detection system," *Applied Sciences*, vol. 13, no. 23, p. 12888, 2023.

[27] S. O'Shaughnessy and S. Sheridan, "Image-based malware classification hybrid framework based on space-filling curves," *Computers & Security*, vol. 116, p. 102660, 2022.

[28] M. M. Belal and D. M. Sundaram, "Global-local attention-based butterfly vision transformer for visualization-based malware classification," *IEEE Access*, 2023.

[29] A. Ravi, V. Chaturvedi, and M. Shafique, "Vit4mal: Lightweight vision transformer for malware detection on edge devices," *ACM Transactions on Embedded Computing Systems*, vol. 22, no. 5s, pp. 1–26, 2023.

[30] D. Maddali, "Convnext-eesnn: An effective deep learning based malware detection in edge based iiot," *Journal of Intelligent & Fuzzy Systems*, no. Preprint, pp. 1–17, 2024.