Privacy preserving and verifiable outsourcing of AI processing for cyber-physical systems

Georgios Spathoulas[1,2][1234−5678−9012], Angeliki Katsika[2][0000−0002−7891−8415], and Georgios Kavallieratos[1][0000−0003−1278−1943]

[1] Dept. of Information Security and Communication Technology, Norwegian University of Science and Technology, Gjøvik, Norway name.surname@ntnu.no
[2] Dept. of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece akatsika@uth.gr

Abstract. Cyber-physical systems (CPSs) have been used in different domains to enable automation, increase efficiency and effectiveness, and reduce the operational costs of traditional systems. CPSs come with several limitations and requirements that must be considered when designing their application to different domains. Artificial intelligence (AI) can facilitate the optimization of cyber-physical systems' operation. However, integration of AI functionality into CPS is not easy due to limitations in hardware, software, and flexibility. The main contribution of the present paper is a novel approach for the use of remote AI services in CPSs and. By employing zero-knowledge proofs, we protect the privacy of models and data and we can verify the integrity of the operations on the side of the AI service. Our experiments have shown that such an approach is feasible and brings significant offerings, such as verifiable remote AI inference for CPS. Our experiments have shown that currently available zero-knowledge implementations require large proof generation times, which hinder the effective application of remote AI services to real-world CPS.

Keywords: Articial Intelligence · Cyber Physical Systems · Cybersecurity · Privacy · Verification

1 Introduction

The unification of embedded systems with communication technologies led to the 'Cyber Physical Systems (CPSs)'. Such systems intertwine physical and cyber components and connect to each other. The increasing proliferation of CPSs in critical domains, including industrial control systems, energy, transportation, and healthcare, increases automation and facilitates operations.

Today, explosive data growth and diverse data types lead to the development of Artificial Intelligence models to refine modeling approaches by improving accuracy and computational capabilities [30]. In recent years, Artificial Intelligence (AI) has been increasingly adopted to control CPS [52]. By leveraging AI in CPSs, more optimized and flexible control is achieved. The great learning and

generalization capabilities of deep neural networks facilitate the deployment of AI models in various domains to handle complex situations in the physical world [32, 40]. AI offers the ability to cognition the system to facilitate its functions and operations. This capability enables the modeling, representation, and learning of complex behaviors and interactions among the CPS components and data. Such capabilities can be achieved through supervised or unsupervised training of AI models designed for these tasks. Additionally, AI models increase the adaptability of CPSs by providing a continuous learning process of the system. Several works examined the integration of AI into CPSs [55]. For example, a distributed approach is used to manage real power generation and consumption schedules by leveraging Artificial Neural Networks [19, 54]. Machine learning algorithms are used in smart grid state estimation to improve the computational efficiency of power systems [16].

Integration of AI in CPSs constitutes a central element of the digital transformation process in any application domain, but is unavoidably accompanied by many challenges. One of those is the enlargement and diversification of the cyber risks that the domain is facing, with existing risks increasing and new risks being introduced. This is mainly due to the fact that whereas traditional operations were designed with no need for cyber security in mind, modern AI - enabled operations are allowed to achieve automated processes and functions. CPSs are time-sensitive systems [2] and hence advanced AI models are needed to ensure the availability of functions and operations. Furthermore, the complexity of CPSs [53] increases the complexity of AI models that need to be deployed. The several components and operations of the CPSs must be considered in the AI models, and the dynamicity and criticality of the environment in which such systems operate must be considered.

The physical and software components of CPSs are intertwined. Several physical devices could share sensitive data with CPS data evaluators to increase the value of their data [59]. Furthermore, data outsourcing facilitates the relocation of data from the CPS for efficient storage at low cost [20]. Although data sharing facilitates several processes and operations in CPSs, data should not be exposed to AI providers and cause privacy leakage. To this end, the privacy of the data should be ensured while retaining the ability to perform the required tasks, such as statistical analysis, classification, and prediction. Recently, there has been increasing attention to an emerging security approach in CPSs known as Zero Trust Architecture (ZTA) [11]. The core principle of the ZTA dictates that no component of the system is trusted by default and every interaction must be authenticated and verified.

Based on the above, modern CPSs tend to adopt AI models to analyze data and facilitate their functions and operations. Contemporary AI models often exhibit considerable size and demand extensive computational resources. Consequently, it is not uncommon for CPS operators to struggle to accommodate such models within the CPS infrastructure itself.

On top of that, it is not feasible for all CPSs operators to collect a qualitative dataset and train an efficient AI model on their side. For cases in which the

operation of the CPS is not highly dependant on the parameters of each different installation, the AI model trained on the dataset collected by a specific CPS installation can be used by all CPS operators.

To this end, the goal of this paper is to identify a methodology that would allow outsourcing AI operations for CPSs in a verifiable and secure way. This will facilitate the analysis of the basic limitations and requirements that such a methodology would create for the operation of the CPS.

The contributions of this work are as follows.

- Identification of how zero knowledge proofs can be employed for providing verifiable AI services to CPS
- Validation of the hypothesis that such an approach would actually preserve data and model privacy and integrity of the inference operations.
- Assessment of computational complexity overhead of such approaches and discussion on their applicability in real-world CPS.

The remainder of this paper is structured as follows. Section 2 reviews the related work and Section 3 presents the background knowledge. Section 4 presents the system reference model. Section 5 presents the application of the proposed approach to the CPS. Finally, Section 6 summarizes our conclusions and outlines the challenges to be addressed in future research.

2 Related work

The applications of AI in CPSs have been extensively studied in the literature. The application of AI in industrial CPSs in seven domains is examined in [52]. The analysis focused on AI controllers and future challenges. A survey is conducted in [8] and the components and interactions of an AI-Augmented industrial CPS are examined. The work emphasized several technological and economic benefits of AI in CPSs, such as real-time monitoring of machinery, efficient maintenance management, and effective scheduling planning. The applications of AI in wireless networking for CPS and Internet of Things are reviewed in [46] focusing on the machine learning paradigms and on the challenges faced by current and future wireless networks related to CPS. The elements of CPSs and AI applications are reviewed in [17] and the need for AI reliability is analyzed. The current and future challenges of the use of AI in CPS are examined in [44] and a taxonomy is provided.

Applications of AI to protect security in CPSs have been extensively studied in the literature [37, 22, 48]. CPSs also face several privacy issues [38, 13]. Different AI techniques to protect privacy in CPSs are investigated in [15]. A differential privacy based scheme for data release in cyber–physical system is proposed in [59]. However, the proposed approach does not utilize AI models. In [36] a Federated Learning approach is proposed for the preservation of data privacy in vehicular CPSs, considering a two-phase mitigating scheme consisting of intelligent data transformation and collaborative data leak detection. In [45], a unified privacy preservation model with AI at the edge is proposed for

human-in-the-loop CPS. The approach focuses on the control and data acquisition supported by AI on the edge.

Several technologies and applications of the ZTA in industrial control systems are reviewed in [11] and their advantages and disadvantages are analyzed. A general set of ZT architectural patterns for CPSs is presented in [18] using the Architecture Analysis and Design Language (AADL) to define the key element of embedded systems. Several ZTA for CPSs in healthcare [10], power IoT [58], and cloud computing [12] are presented in the literature. Several works in the literature leverage AI models to analyze CPSs data. In [29] a federated deep learning scheme is proposed to detect cyber threats against industrial CPSs based on an AI model. An identity-based proxy-authorized outsourcing with public auditing is proposed [1] based on proxy authorization and verification. In [35] an information-centric networking (ICN)-based system model is proposed for CPS that facilitates the processing of data from IIoT devices closer to the edge based on the edge-assisted authentication scheme in CPS.

To our knowledge, the validation and verification of the output of the AI models in CPSs has not yet been explored. Although there are approaches for AI applications in CPSs, ZTA in CPS, and data outsourcing models and frameworks, an approach that ensures the trustworthiness of the outcome of remote AI services in CPSs is needed. The concept of verifiable AI processing is rather new and its application to CPS has not been explored and tested. Such an approach would enable an easier, more flexible and more effective integration of AI technology into CPSs. This paper aims to establish the initial idea, identify the main offerings, research its applicability, and recognize the main challenges that will arise. Since no other relevant work exists, this work can be used as the starting point for other researchers to further advance the proposed approach.

3 Background

3.1 Artificial intelligence in CPSs

The applications of AI in several domains is prominent [24]. Business, finance, healthcare, agriculture, smart cities, cybersecurity, and many more are examples of AI application areas [47]. In recent years, the proliferation of big data through advanced technologies and developments, such as Internet-of-Things (IoT), has spurred the rapid advancement of information retrieval and analysis methodologies, especially artificial intelligence. This progress in handling vast datasets is poised to transform numerous sectors within Industry 4.0. It catalyzes the emergence of smart technologies, where intelligent and automated processes characterize contemporary systems [26]. Additionally, large AI models, or foundation models, are models recently emerging with massive scales both parameter-wise and data-wise, the impacts of which can reach beyond billions [43].

Integration of AI into CPS involves the incorporation of various machine learning algorithms [41], deep learning [29], and reinforcement learning [33] to facilitate decision-making, prediction, and control tasks. Machine learning algorithms, such as neural networks, are used for predictive maintenance, fault

detection, and anomaly detection in CPS [61, 25]. Deep learning techniques, including convolutional neural networks and recurrent neural networks, are utilized for sensor data analysis, image recognition, and natural language processing in CPS applications [39]. Reinforcement learning algorithms enable CPS to learn optimal control policies and adapt to dynamic environments through interaction and feedback [42]. This integration in CPS facilitates the adoption of autonomous and data-driven solutions to support the main functions and operations of the CPSs. AI approaches in CPSs have been proposed in several critical sectors such as in healthcare [31], smart logistics [34], and smart manufacturing [27].

3.2 Zero Knowledge

The notion of zero-knowledge represents a significant advancement in secure communication and authentication. Zero-Knowledge Proofs (ZKPs) operate on a fundamental principle: A prover can demonstrate the truth of a statement to a verifier without revealing any additional information beyond the truthfulness of the statement. This approach prioritizes privacy and security, ensuring that sensitive data are protected while undergoing verification processes.

The inception of ZKPs dates back to the seminal work of Goldwasser, Rackoff, and Micali [14], while in recent years zero-knowledge proofs have gained significant interest, as they have been proven crucial in scaling compute-constrained networks and enhancing user privacy in blockchain technology. Beyond blockchains, they facilitate efficient verification of computations, extending trust and verifiability to machine learning models and various real-world applications.

To ensure effectiveness and reliability in practical applications, every zero-knowledge proof protocol adheres to the following foundational principles:

- **Completeness** ensures that if the input provided to the ZKP is valid, and both the prover and verifier act honestly, the proof is accepted without fail, as the protocol consistently returns 'true'.
- **Soundness** dictates that a dishonest prover cannot deceive an honest verifier into accepting an invalid statement as true. If the input to the protocol is invalid, it should be theoretically impossible for a dishonest prover to manipulate the protocol into returning 'true'.
- **Zero-Knowledge** ensures that the verifier does not obtain additional information about the statement beyond its validity, preventing any insight into its contents or the derivation from the proof.

Zero-knowledge proofs generally encode programs as arithmetic circuits. Through these circuits, the prover generates a proof based on both public and private inputs, while the verifier mathematically verifies the correctness of the output statement without accessing any details about the private inputs.

Zero-knowledge systems are built upon the most widely adopted protocols, namely ZK-SNARKs and ZK-STARKs:

- **ZK-SNARK**, or Zero-Knowledge Succinct Non-Interactive Argument of Knowledge [4], [5], provides a significant cryptographic tool allowing one

party to verify a computation's validity without revealing its inputs. This eliminates the need for the verifier to execute the computation and enables short and succinct proofs compared to other cryptographic protocols [56]. However, the protocol relies on a trusted setup ceremony, where initial parameters are generated in a secure environment.
- **ZK-STARK** which stands for zero-knowledge Scalable Transparent Argument of Knowledge [3], serves as an alternative to SNARKs, eliminating the need for a trusted setup and utilizing hash functions. Despite offering benefits such as quantum resistance and eliminating the need for a trusted setup, zk-STARKs come with larger proof sizes, resulting in longer verification times depending on the implementation.

Research in ZKP protocols has led to advances that resulted in smaller memory footprints and faster processing times. These developments have facilitated the verification of machine learning algorithms on-chain, marking a significant advancement in the integration of ZKP with Machine Learning (ML). The applications of zero-knowledge proofs in machine learning can be categorized into distinct domains, each serving specific functionalities and purposes, encompassing model verification, data privacy, and validation processes:

- **Model Verification and Authenticity:** ZKP can be used a) to ensure the validity and integrity of machine learning models, preventing instances where a different model is served than claimed, and b) to verify the consistent application of machine learning algorithms across different users' data.
- **Data Privacy and Security:** ZKP can be applied a) to preserve data privacy and confidentiality during decentralized machine learning inference or training, and additionally, b) to facilitate proof of personhood, verifying individuals without disclosing identifiable information, which is essential for privacy and sybil-resistance.

Achieving a correct execution proof for machine learning models in zero-knowledge systems requires encoding various model elements like architecture, parameters, constraints, and operations into arithmetic circuits. However, this field is still developing, and while promising optimizations like proof recursion and emerging frameworks are on the horizon, the challenge remains to align zero-knowledge proof with the increasing complexity of machine learning models. In the following paragraph, we present some notable work on the integration of zero-knowledge techniques with machine learning.

In [28] Lee et al. devised an efficient verifiable Convolutional Neural Network (vCNN) framework to improve proving performance using a novel relation representation for convolution equations, reducing proving complexity from $O(ln)$ in existing zk-SNARK approaches to $O(l + n)$ in their proposed method, where l and n denote the size of the kernel and the data in CNNs. In another approach [57], Weng et al. proposed a method for privacy-preserving and verifiable CNNs, employing a QMP (Quadratic Matrix Program)-based arithmetic circuit to represent convolutional relations and generate zkSNARKs proofs using Homomorphic Encryption (HE) and collaborative inference, achieving in their

experiments faster setup and proving time, compared to the QAP(Quadratic Arithmetic Program)-based method. Camuto and Morton developed the EZKL framework[6], as a command-line tool to build zkSNARKs specific to the inference phase of machine-learning models. EZKL uses Halo2 with KZG in the back-end, making it compatible with larger models compared to other zkSNARK implementations. Kang et al.[23] also created Halo2 proofs to authenticate the inference phase of the MLaaS (ML-as-a-service) model. They introduced an ImageNet-scale zkSNARK that uses Halo2[60] and outlined three protocols to verify the accuracy, predictions, and trustless retrieval of the ML model.

4 System

4.1 Reference architecture

To the end of presenting our approach, we define in this subsection a typical set-up that can be viewed as a reference scenario to which the proposed approach can be applied to.

The main actors of such a scenario are the following:

- The CPS operator
- The AI service provider
- The training data provider

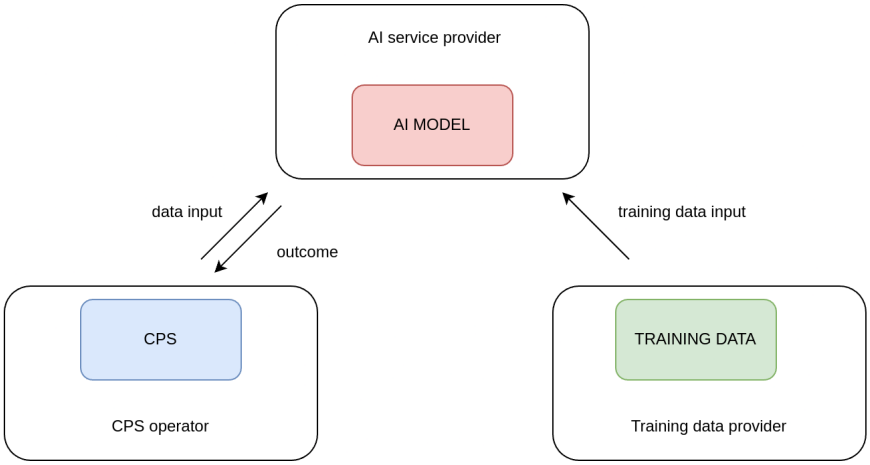

Fig. 1. Reference setup

The main interactions between the components of these actors are shown in Figure 1. The CPS produces series of measurement values that have to be fed to an AI model in order to produce a meaningful outcome or to trigger

specific actions for the CPS's actuators. In a simple setup, this could all happen internally in the CPS, but due to a number of restrictions, this is usually not feasible. Such restrictions are related to the following claims:

- Modern AI models tend to be extremely large and require a lot of computational resources. Due to this, it is common for CPS operators to be unable to support such models in the CPS itself.
- AI models may need to be periodically updated under reinforcement learning schemes, and the installation of CPS may not offer flexibility with respect to this requirement.
- The access to the AI model itself may be restricted. Today, AI models tend to have closed access and are offered as a service (free or paid) to others.
- Data for model training may not be available to the CPS installation owner. AI models require voluminous and qualitative labeled datasets, and the production of those may not be feasible at each CPS installation.

A setup that enables overcoming all previous points is to use an externally provided AI trained model. In this approach, the AI model is managed and offered as a service by another actor indicated as the AI service provider. The outsourcing of the inference step would enable the overcoming of the mentioned limitations but at the same time it would create concerns related to the integrity of the operations that happen on the AI service provider's side.

The main concept behind our methodology is to support this outsourcing operation with zero-knowledge proofs, and thus resolve any integrity issues and supporting model privacy. The CPS system sends a data point or a series of data points to the AI service provider, which is fed to the AI model, and the result produced is returned to the CPS. The AI service provider has to commit to a specific AI model, though a cryptographic commitment to it (based on ZK proofs) which is shared with the CPSs. The AI service provider can then produce a validity proof (based on ZK proofs) of the operations of the AI model during the inference step and provide this proof to the CPS. The latter can then validate that the operations have been performed according to the appropriate agreed AI model. In addition, the fact that the AI model remains with the AI service provider protects the privacy of the training data. Multiple attacks[7, 9] have been reported with regard to the extraction of training data from trained models.

4.2 Components

On the side of the CPS operator, there are several sensors and several actuators. In the generic case, the values detected by the sensors can be used to identify the optimal action to be assigned to the actuators. In a system with n sensors $s_1, s_2, ...s_n$ and m actuators $a_1, a_2, ...a_m$, the optimal action for each actuator at time t can be defined as a function of the recent values measured by the CPS sensors.

$$action_t^i = f(s_{t-1}^1..., s_{t-r}^1, ..., s_{t-1}^n..., s_{t-r}^n) \quad (1)$$

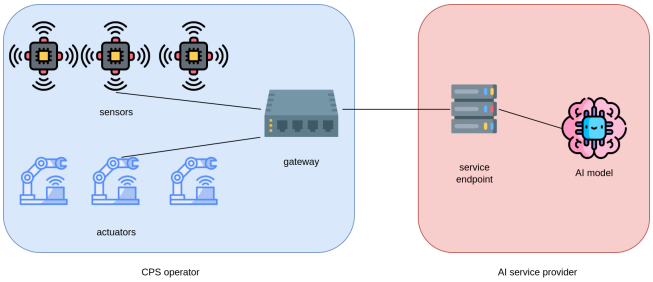

Fig. 2. Main components

The r most recent measurements of all sensors are taken as input to a specific processing process to define the next action of the actuator i. If this processing process is simple, straightforward, and lightweight in computations, then it can be integrated into the CPS. If that processing is more complex (e.g. the input is fed to a ML model), then such integration becomes difficult. Especially in the case of applying ML models to the sensors output to decide the actuators' decisions, the restrictions mentioned in the previous subsection come into play.

In the present paper, we propose an architecture that can be applied to such cases and allows employing ML models offered by other actors to operate CPS.

In the context of the previously analyzed setup, the CPS operator needs to support the system by providing a gateway component that will collect all data detected by the sensors and send it to the AI service provider. The gateway component then receives the data output from the AI service provider and uses that to operate the CPS's actuators.

On the side of the AI service provider, the main components are the AI trained model and the service endpoint, which manages the communications with the CPS operator(s). The service will be accessed by more than one CPS operator at the same time, so the service endpoint also manages concurrent connections.

4.3 Operation

The operation of the system mainly comprises two phases, the commitment generation stage which occurs initially only once for each instance of the AI model used and the operation phase which is repeated for each used of the AI model by the CPS.

Commitment generation Let us assume that the AI service provider starts with an untrained AI model denoted as m_0. Based on the training data that the AI provider maintains or the training data that other actors may make available, the initial model m_0 is trained and transformed to a new instance m_1.

$$m_1 = training(m_0, data) \tag{2}$$

The newly trained instance can be used by the CPS, and the AI service provider must commit to that before doing so. The commitment generation process takes as input the model m_1 and produces a cryptographic commitment c_1.

$$c_1 = commit(m_1) \tag{3}$$

Commitment generation has to be repeated in any subsequent update of the model. For every new training phase that may occur for the model:

$$m_i = training(m_{i-1}, data) \tag{4}$$

the commitment generation process is required:

$$c_i = commit(m_i) \tag{5}$$

The commitment produced is directly connected to the trained instance of the model upon which it has been produced. This commitment has to be shared with potential CPSs that will make use of the model in a secure way. The integrity of the commitment data is crucial for the general workflow.

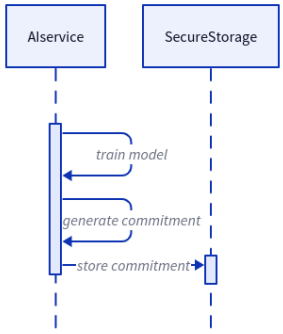

Fig. 3. Commitment generation

Verifiable inference Consequently, and given the fact that the CPS has access to the commitment c_i produced for the model m_i that the AI service provider offers as a service, the CPS can make use of the service in a verifiable way.

We assume that the AI model maintained by the AI service provider can decide the optimal set of actions for the CPS's actuators at a specific time point t:

$$actions_t = [action_t^1, ..., action_t^m] \tag{6}$$

The input required by the AI model to produce this output is the r most recent values monitored by all n CPS sensors and is denoted as $values_{t,r}$.

$$values_{t,r} = \begin{bmatrix} s_{t-1}^1 & s_{t-2}^1 & ... & s_{t-r}^1 \\ s_{t-1}^2 & s_{t-2}^2 & ... & s_{t-r}^2 \\ ... & ... & ... & ... \\ s_{t-1}^n & s_{t-2}^n & ... & s_{t-r}^n \end{bmatrix} \tag{7}$$

The process takes place in the following steps :

- In the first step of the inference phase, the CPS system sends to the AI service provider $values_{t,r}$ in a private encrypted format.
- The AI service provider feeds the data to the AI model m_i and calculates the corresponding output that relates to the best actions for the CPS actuators $actions_t$.
- The AI service provider generates a proof related to the previous-step operations.

$$proof_t = proofgen(m_i, values_{t,r}, actions_t) \tag{8}$$

- The AI service operator returns to the CPS the $actions_t$ along with $proof_t$.
- The CPS verifies the validity of the operations of the AI service provider. It used as input the received data $actions_t$ and $proof_t$ in combination with the sent data $values_{t,r}$ and the commitment c_i received in the commitment generation phase.

$$result = proofval(c_i, proof_t, values_{t,r}, actions_t) \tag{9}$$

- If the result of the verification in the previous step is positive, then the CPS can apply the received $actions_t$ to the actuators.

5 Experiments

Although the proposed approach is promising and may change the way AI is integrated in CPSs, the field of zero-knowledge is still under heavy research and development, and existing implementations tend to be restricted in terms of efficiency and size of processing that can be verified. In the present Section, we run a series of experiments to evaluate if presently offered implementations of zero-knowledge machine learning can serve the needs of CPSs.

The experiments are based on the concept that the operator of an industrial control system (ICS) needs to monitor the operation of the installation. The installation includes steam-turbine power generation and pumped-storage

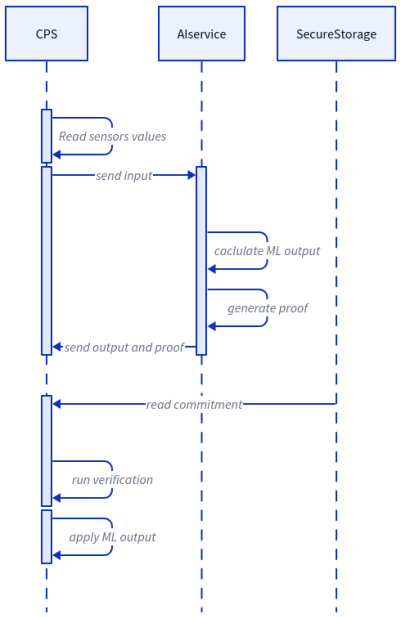

Fig. 4. Verifiable inference

hydropower generation subsystems. There is an AI service provider that offers an AI model that can process monitoring data for the operation of the installation and conclude on the existence of a cybersecurity attack. The CPS gathers monitoring data and sends that to the service provider to perform the inference step for the offered AI model following the methodology described in Section 4. The goal of the experiment is twofold; to validate that it is feasible to apply the proposed methodology, and to assess the processing overhead that the proposed methodology brings in relation to the size of the AI model used.

5.1 Dataset

To do this, we used the HAI (HIL-based Augmented ICS) Security Dataset [51, 50, 49]. The HAI dataset was collected from a realistic industrial control system (ICS) testbed augmented with a Hardware-In-the-Loop (HIL) simulator that emulates steam-turbine power generation and pumped-storage hydropower generation.

The testbed consists of four different processes: boiler process, turbine process, water treatment process, and HIL simulation:

- Boiler Process (P1): This includes a water-to-water heat treater at low pressure and moderate temperature. This process is controlled using Emerson Ovation DCS.
- Turbine Process (P2): A rotor kit process that closely simulates the behavior of an actual rotating machine. It is controlled by GE's Mark VIe DCS.
- Water Treatment Process (P3): This process includes pumping water to the upper reservoir and releasing it back into the lower reservoir. It is controlled by Siemens's S7-300 PLC.
- HIL Simulation(P4): Both the boiler and turbine processes are interconnected to synchronize with the rotating speed of the virtual steam-turbine power generation model. The pump and value in the water-treatment process are controlled by the pumped-storage hydropower generation model.

During simulation a number of attacks have been conducted based on combinations of the following basic attacker steps:

- Process Variables (PV) response prevention: An attacker can hide their attack by covering up the PV response because PV is the fundamental measurement to monitor current operating condition.
- Set Point (SP) attack: An attacker can change the SP and then naturally manipulate the PV as desired. The controller automatically adjusts the CO until the relevant PV reaches the SP when an operator changes the set point.
- Control Output (CO) attack: An attacker can directly control the actuators by changing the CO values. This attack can cause actuator malfunctions and disrupt process production.

The data set consists of 84 fields that include the timestamp of the measurements, 79 measurement values from different sensors in the system, and 4 labels (3 related to the attack category and 1 relating to the existence of any attack).

5.2 Experiments setup

Regarding zero-knowledge machine learning, we have chosen to use EZKL[6], which can take a high-level description of a program and set up a zero-knowledge prover and verifier. EZKL focuses on programs that are expressed as pytorch[21] AI/ML models and other computational graphs. After setup, the prover can prove statements such as the following. These proofs can be trusted by anyone with a copy of the verifier and verified locally or even directly on Ethereum and compatible chains.

We aim to simulate the operation of a remote service that could offer AI functionality to the CPS of the dataset[51, 50, 49]. The CPS operator sends the data sensed by the sensors to the AI service provider, and the result is related to whether there is an abnormality in the measurements (relevant to any possible attack). At each run experiment, we trained a neural network of a specific size, with the provided data, and tested the validity proof generation and validation process along with the time required for proof generation, which is the most time-consuming process of the workflow. As shown in Figure 5, each neural network

has three layers; the input layers consisting of 79 nodes (that correspond to the 79 different sensors at a time point), a hidden layer whose size was varied from experiment to experiment and the output layer which has only one node that makes the decision on the existence of an abnormality at that specific time point.

We varied the size of the hidden layer, to create different size neural networks and test how the proof generation process is related to that size. Specifically, the size of the hidden layer has been set to 20,40,80 and 100 nodes. We used a VM with 8GBs of RAM and one vCPU on top of an Intel(R) Xeon(R) Silver 4210 CPU 2.20GHz.

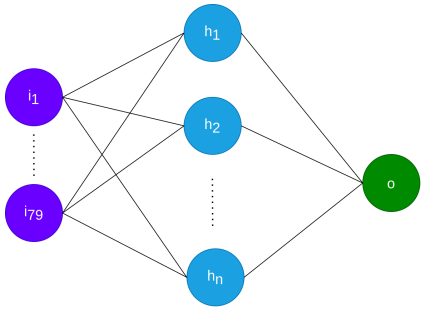

Fig. 5. Neural networks

5.3 Results

The experiments run smoothly. The EZKL library performed as expected and can enable the CPS operator to use a remote AI model in a privacy-enabling and verifiable way. We verify that the operation is as expected and that

- Any change in the data input or in the model that the AI service provider uses ends up with a proof that cannot be verified.
- The CPS operator has no access to the ML model itself or its parameters.

For each of the experiments with neural networks of different sizes we measured both the time required for the proof generation on the side of the AI service provider and the time required for the proof verification on the side of the CPS. Because the time is highly dependent on the values of randomly picked cryptographic parameters, we observed that running the same experiment (with same size neural networks) twice can produce significantly different measured time values. For that reason, we opted for running each experiment 20 times and providing an aggregation of the time values measured.

The results of the proof generation time are shown in Figure 6 and the results of the proof verification time are depicted in Figure 7.

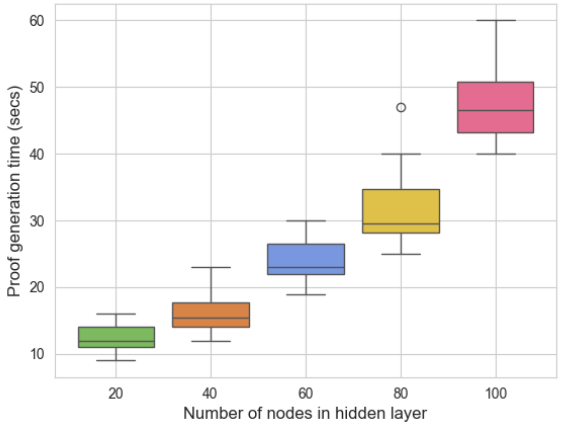

Fig. 6. Proof generation time

As is evident from Figure 6, the time required to generate the proof of properly processing a single data input instance is significantly high. It starts at approximately 12 seconds on average for the network with the 20 nodes in the hidden layer and is gradually increased in subsequent experiments. It reaches an average of 45 seconds for the case of a network with 100 nodes in the hidden layer.

Figure 7 shows that the time required for verification is significantly lower. It starts around 200 msecs for the smaller neural networks and reaches up to 800 msecs for the network with 100 nodes in the hidden layer.

The accuracy of the tested models varies according to the size of the hidden layer. Smaller networks have lower accuracy in detecting attacks, but the experiments have shown that the networks with a hidden layer of 60 nodes and upwards achieve an accuracy of at least 90% which is satisfactory for the given classification problem.

Before discussing the time results, we must indicate that the results obtained have been achieved on minimal hardware resources (a single vCPU). The use of more capable hardware and probably the use of GPUs to efficiently produce the required proofs will certainly decrease the required time by at least an order of magnitude. Even in that case, it is understandable that using a remote verifiable AI service for real-time decisions/actions in the CPS is currently infeasible. However, there are cases for which the delay imposed by the proof generation step is not so critical. For example, if the CPS installation needs to periodically check its status every 60 seconds to decide if there is any evidence of malfunction or even security breach, the delay imposed by the proof generation step is bearable.

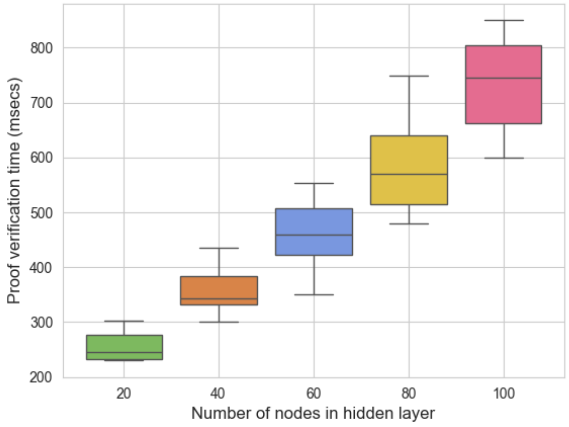

Fig. 7. Proof verification time

The delay imposed by the verification step is minimal and can be handled without employing significant hardware resources on the side of the client.

6 Discussion

In the present paper we tested the application of verifiable processing through the use of emerging research results in the zero-knowledge field in the CPSs domain. The main findings of our research are summarized in the present Section.

The use of such an approach for CPS indicates a very interesting research direction, as the offerings are highly relevant to the requirements of the domain. Providing remote AI services in the way discussed has the following advantages:

- Minimises the hardware/software requirements in the CPS itself
- Allows for new and update-able AI systems to be applied to a deployed CPS without any modifications to it.
- Protects privacy of AI models, and training data.
- Allows for the use of such services in critical operations as the methodology provides cryptographic verification of the validity of the output returned by the AI service.
- The methodology does not add additional hardware requirements for verification.

On the other hand, there are a number of limitations which cannot be overlooked. The main limitation relates to the high proof generation time, which depends on the size of the AI model. The aforementioned limitation creates a costly requirement for high-efficiency hardware to be used on the side of the AI service provider. Even if such hardware may already be available to the AI

service provider, the allocation of it to proof generation instead of AI model training creates additional cost of operation.

Nevertheless, the application of zero-knowledge proofs to allow for verifiable remote use of AI models is highly promising. As plans for future work, we aim at exploring the use of GPU hardware to assess to what extent can such an option provide a performance boost and a decrease of the proof generation time. Additionally, other implementations of zero-knowledge machine learning apart from EZKL are being developed, and testing those would also be an interesting future direction for our research. On top of that, we will study more thoroughly the CPSs uses cases in which this approach could be used and what the specific requirements stemming from those use cases are. The integration of such an approach into real-world CPS would reveal additional parameters and requirements that have to be taken into account when developing the scheme.

Funding: This work was supported by the European Commission [grant 101120657 "European Lighthouse to Manifest Trustworthy and Green AI" - ENFIELD].

References

1. Al-Turjman, F., Deebak, B.: A proxy-authorized public auditing scheme for cyber-medical systems using ai-iot. IEEE Transactions on Industrial Informatics **18**(8), 5371–5382 (2021)
2. Alcaraz, C., Lopez, J.: Analysis of requirements for critical control systems. International journal of critical infrastructure protection **5**(3-4), 137–145 (2012)
3. Ben-Sasson, E., Bentov, I., Horesh, Y., Riabzev, M.: Scalable, transparent, and post-quantum secure computational integrity. Cryptology ePrint Archive, Paper 2018/046 (2018), https://eprint.iacr.org/2018/046, https://eprint.iacr.org/2018/046
4. Bitansky, N., Canetti, R., Chiesa, A., Tromer, E.: From extractable collision resistance to succinct non-interactive arguments of knowledge, and back again. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. p. 326–349. ITCS '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2090236.2090263, https://doi.org/10.1145/2090236.2090263
5. Bitansky, N., Chiesa, A., Ishai, Y., Paneth, O., Ostrovsky, R.: Succinct non-interactive arguments via linear interactive proofs. In: Sahai, A. (ed.) Theory of Cryptography. pp. 315–333. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
6. Camuto, A.D., Morton, J.: Ezkl (2024), https://github.com/zkonduit/ezkl
7. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 5253–5270 (2023)
8. Chae, J., Lee, S., Jang, J., Hong, S., Park, K.J.: A survey and perspective on industrial cyber-physical systems (icps): From icps to ai-augmented icps. IEEE Transactions on Industrial Cyber-Physical Systems (2023)
9. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., Mukhopadhyay, D.: A survey on adversarial attacks and defences. CAAI Transactions on Intelligence Technology **6**(1), 25–45 (2021)

10. Chen, B., Qiao, S., Zhao, J., Liu, D., Shi, X., Lyu, M., Chen, H., Lu, H., Zhai, Y.: A security awareness and protection system for 5g smart healthcare based on zero-trust architecture. IEEE Internet of Things Journal **8**(13), 10248–10263 (2020)
11. Feng, X., Hu, S.: Cyber-physical zero trust architecture for industrial cyber-physical systems. IEEE Transactions on Industrial Cyber-Physical Systems **1**, 394–405 (2023)
12. Ferretti, L., Magnanini, F., Andreolini, M., Colajanni, M.: Survivable zero trust for cloud computing environments. Computers & Security **110**, 102419 (2021)
13. Giraldo, J., Sarkar, E., Cardenas, A.A., Maniatakos, M., Kantarcioglu, M.: Security and privacy in cyber-physical systems: A survey of surveys. IEEE Design & Test **34**(4), 7–17 (2017)
14. Goldwasser, S., Micali, S., Rackoff, C.: The knowledge complexity of interactive proof-systems. In: Symposium on the Theory of Computing (1985), https://api.semanticscholar.org/CorpusID:209402113
15. Gupta, R., Tanwar, S., Al-Turjman, F., Italiya, P., Nauman, A., Kim, S.W.: Smart contract privacy protection using ai in cyber-physical systems: tools, techniques and challenges. IEEE access **8**, 24746–24772 (2020)
16. Hadayeghparast, S., Karimipour, H.: Application of machine learning in state estimation of smart cyber-physical grid. Security of Cyber-Physical Systems: Vulnerability and Impact pp. 169–194 (2020)
17. Haldorai, A.: A review on artificial intelligence in internet of things and cyber physical systems. Journal of Computing and Natural Science **3**(1), 012–023 (2023)
18. Hasan, S., Amundson, I., Hardin, D.: Zero trust architecture patterns for cyber-physical systems. Tech. rep., SAE Technical Paper (2023)
19. Hinrichs, C., Lehnhoff, S., Sonnenschein, M.: Cohda: A combinatorial optimization heuristic for distributed agents. In: Agents and Artificial Intelligence: 5th International Conference, ICAART 2013, Barcelona, Spain, February 15-18, 2013. Revised Selected Papers 5. pp. 23–39. Springer (2014)
20. Huang, K., Zhang, X., Wang, X., Mu, Y., Rezaeibagha, F., Xu, G., Wang, H., Zheng, X., Yang, G., Xia, Q., et al.: Hucdo: A hybrid user-centric data outsourcing scheme. ACM Transactions on Cyber-Physical Systems **4**(3), 1–23 (2020)
21. Imambi, S., Prakash, K.B., Kanagachidambaresan, G.: Pytorch. Programming with TensorFlow: Solution for Edge Computing Applications pp. 87–104 (2021)
22. Jamal, A.A., Majid, A.A.M., Konev, A., Kosachenko, T., Shelupanov, A.: A review on security analysis of cyber physical systems using machine learning. Materials Today: Proceedings **80**, 2302–2306 (2023)
23. Kang, D., Hashimoto, T., Stoica, I., Sun, Y.: Scaling up trustless dnn inference with zero-knowledge proofs (2022)
24. Kasula, B.Y.: Advancements and applications of artificial intelligence: A comprehensive review. International Journal of Statistical Computation and Simulation **8**(1), 1–7 (2016)
25. Kim, S., Park, K.J.: A survey on machine-learning based security design for cyber-physical systems. Applied Sciences **11**(12), 5458 (2021)
26. Kim, S.W., Kong, J.H., Lee, S.W., Lee, S.: Recent advances of artificial intelligence in manufacturing industrial sectors: A review. International Journal of Precision Engineering and Manufacturing pp. 1–19 (2022)
27. Lee, J., Li, W., Hsu, Y.M., Jia, X.: Cyber–physical systems framework for ai in smart manufacturing and maintenance. In: Artificial Intelligence in Manufacturing, pp. 233–272. Elsevier (2024)

28. Lee, S., Ko, H., Kim, J., Oh, H.: vcnn: Verifiable convolutional neural network based on zk-snarks. IEEE Transactions on Dependable and Secure Computing pp. 1–17 (2023). https://doi.org/10.1109/TDSC.2023.3348760
29. Li, B., Wu, Y., Song, J., Lu, R., Li, T., Zhao, L.: Deepfed: Federated deep learning for intrusion detection in industrial cyber–physical systems. IEEE Transactions on Industrial Informatics **17**(8), 5615–5624 (2020)
30. Li, J., Herdem, M.S., Nathwani, J., Wen, J.Z.: Methods and applications for artificial intelligence, big data, internet of things, and blockchain in smart energy management. Energy and AI **11**, 100208 (2023)
31. Liu, W., Zhao, F., Shankar, A., Maple, C., Peter, J.D., Kim, B.G., Slowik, A., Parameshachari, B., Lv, J.: Explainable ai for medical image analysis in medical cyber-physical systems: Enhancing transparency and trustworthiness of iomt. IEEE Journal of Biomedical and Health Informatics (2023)
32. Liu, W., Mehdipour, N., Belta, C.: Recurrent neural network controllers for signal temporal logic specifications subject to safety constraints. IEEE Control Systems Letters **6**, 91–96 (2021)
33. Liu, X., Xu, H., Liao, W., Yu, W.: Reinforcement learning for cyber-physical systems. In: 2019 IEEE International Conference on Industrial Internet (ICII). pp. 318–327 (2019). https://doi.org/10.1109/ICII.2019.00063
34. Liu, Y., Tao, X., Li, X., Colombo, A.W., Hu, S.: Artificial intelligence in smart logistics cyber-physical systems: State-of-the-arts and potential applications. IEEE Transactions on industrial cyber-physical systems **1**, 1–20 (2023)
35. Lu, Y., Wang, D., Obaidat, M.S., Vijayakumar, P.: Edge-assisted intelligent device authentication in cyber–physical systems. IEEE Internet of Things Journal **10**(4), 3057–3070 (2022)
36. Lu, Y., Huang, X., Dai, Y., Maharjan, S., Zhang, Y.: Federated learning for data privacy preservation in vehicular cyber-physical systems. IEEE Network **34**(3), 50–56 (2020)
37. Lv, Z., Chen, D., Lou, R., Alazab, A.: Artificial intelligence for securing industrial-based cyber–physical systems. Future generation computer systems **117**, 291–298 (2021)
38. Nazarenko, A.A., Safdar, G.A.: Survey on security and privacy issues in cyber physical systems. AIMS Electronics and Electrical Engineering **3**(2), 111–143 (2019)
39. Ni, P., Li, Y., Li, G., Chang, V.: A hybrid siamese neural network for natural language inference in cyber-physical systems. ACM Trans. Internet Technol. **21**(2) (mar 2021). https://doi.org/10.1145/3418208, https://doi.org/10.1145/3418208
40. Nivison, S.A., Khargonekar, P.P.: Development of a robust deep recurrent neural network controller for flight applications. In: 2017 American Control Conference (ACC). pp. 5336–5342. IEEE (2017)
41. Olowononi, F.O., Rawat, D.B., Liu, C.: Resilient machine learning for networked cyber physical systems: A survey for machine learning security to securing machine learning for cps. IEEE Communications Surveys & Tutorials **23**(1), 524–552 (2021). https://doi.org/10.1109/COMST.2020.3036778
42. Padakandla, S.: A survey of reinforcement learning algorithms for dynamically varying environments. ACM Computing Surveys (CSUR) **54**(6), 1–25 (2021)
43. Qiu, J., Li, L., Sun, J., Peng, J., Shi, P., Zhang, R., Dong, Y., Lam, K., Lo, F.P.W., Xiao, B., et al.: Large ai models in health informatics: Applications, challenges, and the future. IEEE Journal of Biomedical and Health Informatics (2023)
44. Radanliev, P., De Roure, D., Van Kleek, M., Santos, O., Ani, U.: Artificial intelligence in cyber physical systems. AI & society **36**, 783–796 (2021)

45. Rivadeneira, J.E., Borges, G.A., Rodrigues, A., Boavida, F., Silva, J.S.: A unified privacy preserving model with ai at the edge for human-in-the-loop cyber-physical systems. Internet of Things **25**, 101034 (2024)
46. Salau, B.A., Rawal, A., Rawat, D.B.: Recent advances in artificial intelligence for wireless internet of things and cyber–physical systems: A comprehensive survey. IEEE Internet of Things Journal **9**(15), 12916–12930 (2022)
47. Sarker, I.H.: Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. SN Computer Science **3**(2), 158 (2022)
48. Sedjelmaci, H., Guenab, F., Senouci, S.M., Moustafa, H., Liu, J., Han, S.: Cyber security based on artificial intelligence for cyber-physical systems. IEEE Network **34**(3), 6–7 (2020)
49. Shin, H.K., Lee, W., Yun, J.H., Kim, H.: HAI 1.0: HIL-Based Augmented ICS Security Dataset. USENIX Association, USA (2020)
50. Shin, H.K., Lee, W., Yun, J.H., Min, B.G.: Two ics security datasets and anomaly detection contest on the hil-based augmented ics testbed. In: Cyber Security Experimentation and Test Workshop. p. 36–40. CSET '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474718.3474719, https://doi.org/10.1145/3474718.3474719
51. Shin, H.K.e.a.: Hai security datasets (2023), https://github.com/icsdataset/hai
52. Song, J., Lyu, D., Zhang, Z., Wang, Z., Zhang, T., Ma, L.: When cyber-physical systems meet ai: a benchmark, an evaluation, and a way forward. In: Proceedings of the 44th International Conference on Software Engineering: Software Engineering in Practice. pp. 343–352 (2022)
53. Spathoulas, G., Kavallieratos, G., Katsikas, S., Baiocco, A.: Attack path analysis and cost-efficient selection of cybersecurity controls for complex cyberphysical systems. In: European Symposium on Research in Computer Security. pp. 74–90. Springer (2021)
54. Veith, E.: Universal smart grid agent for distributed power generation management. Logos Verlag Berlin (2017)
55. Veith, E.M., Fischer, L., Tröschel, M., Nieße, A.: Analyzing cyber-physical systems from the perspective of artificial intelligence. In: Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control. pp. 85–95 (2019)
56. Wahby, R.S., Tzialla, I., Shelat, A., Thaler, J., Walfish, M.: Doubly-efficient zksnarks without trusted setup. In: 2018 IEEE Symposium on Security and Privacy (SP). pp. 926–943 (2018). https://doi.org/10.1109/SP.2018.00060
57. Weng, J., Weng, J., Tang, G., Yang, A., Li, M., Liu, J.N.: pvcnn: Privacy-preserving and verifiable convolutional neural network testing (2023)
58. Xiaojian, Z., Liandong, C., Jie, F., Xiangqun, W., Qi, W.: Power iot security protection architecture based on zero trust framework. In: 2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP). pp. 166–170. IEEE (2021)
59. Ye, H., Liu, J., Wang, W., Li, P., Li, T., Li, J.: Secure and efficient outsourcing differential privacy data release scheme in cyber–physical system. Future Generation Computer Systems **108**, 1314–1323 (2020)
60. Zcash: Halo2 (2024), https://zcash.github.io/halo2/
61. Zhang, J., Pan, L., Han, Q.L., Chen, C., Wen, S., Xiang, Y.: Deep learning based attack detection for cyber-physical system cybersecurity: A survey. IEEE/CAA Journal of Automatica Sinica **9**(3), 377–391 (2021)