IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 February 2025, accepted 5 March 2025, date of publication 17 March 2025, date of current version 27 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3552112

RESEARCH ARTICLE

AI-Based Detection and Simulation of Hardware Trojan Attacks in Industrial Applications

ALESSANDRO MASSARO[1], (Senior Member, IEEE), NICOLA EPICOCO[1], (Member, IEEE), GIUSEPPE LOSETO[1], (Member, IEEE), AND CARMELO A. ARDITO[2]

[1]Department of Engineering, LUM "Giuseppe Degennaro" University, 70010 Casamassima, Italy
[2]Department of Electrical and Information Engineering, Polytechnic University of Bari, 70125 Bari, Italy

Corresponding author: Nicola Epicoco (epicoco@lum.it)

This work was supported by the Framework of the Project "xTech NextHub: Competence Center for Innovative Solutions Development" through the Bando Regionale della Puglia per gli Aiuti in Esenzione 17 del 30/09/2014-BURP 139 suppl. del 06/10/2014 e s.m.i.—Titolo II Capo 2 del Regolamento Generale "Avviso per la Presentazione dei Progetti Promossi da Grandi Imprese ai Sensi Dell'articolo 17 del Regolamento."

ABSTRACT Industrial Internet of Things (IIoT) is becoming increasingly important for companies worldwide since it allows integrating data from different areas and sectors, thus creating new opportunities to raise efficiency, flexibility, and productivity. However, IIoT also exposes companies to new risks and threats, which implies new challenges related to cybersecurity. In this context, this paper analyzes Hardware Trojan (HT) attacks also characterized by possible malicious controlled white noises on sensors working in an IIoT scenario. Through a parametric analysis in a simulation environment, the impact and the sensitivity of the hardware attacks about the electrical signal deterioration is analyzed as a function of the variation of the HT resistance and capacitance, also superposing a white noise effect to increase the signal disturbance and complexity to find possible causes. The model also includes the possibility for the attackers to modulate and switch the white noise signal in accordance with the IIoT transmission triggering. Specifically, the paper analyzes the behavior of an attacked resistive humidity sensor by means of circuit simulations. Furthermore, the paper proposes a new approach for training commonly used Artificial Intelligence (AI) algorithms, such as Artificial Neural Networks (ANNs) and Random Forest (RF), to predict disturbed sensor signals. Simulation results show the proposed method is effective in detecting HT attacks, thus supporting the adoption of proper corrective actions. The AI-based data processing also demonstrates that circuit simulations provide an effective solution to train AI models tailored to a digital twin HT cybersecurity framework.

INDEX TERMS Hardware trojan, hardware cybersecurity, artificial neural networks, noise prediction.

I. INTRODUCTION

Industry 4.0 represents a transformative shift in manufacturing and industrial practices evolving traditional industries into highly automated and interconnected systems [1]. It enables real-time monitoring, predictive maintenance, and the seamless flow of information across the supply chain. In particular, Industry 4.0 integrates advanced technologies, such as robotic automation, cloud computing, Industrial Internet of Things (IIoT), and Artificial Intelligence (AI), into several industrial processes (see, e.g., the contributions in [2], [3], and [4]). It also emphasizes the importance of data-driven decision-making, allowing businesses to enhance operational performance, reduce costs, and create more customized products [5]. Nowadays, due to the continuous exchange of a large amount of data in and between companies [6], [7], cybersecurity plays a pivotal role to protect the company productivity [8], thus becomining a relevant challenge for all companies worldwide [9].

Usually, cyberattacks occur through the internet network, targeting various aspects of the digital infrastructure of companys, including servers, databases, communication

The associate editor coordinating the review of this manuscript and approving it for publication was Stefano Scanzio.

systems, and any IIoT device connected to the network [10]. The effects of such attacks can be catastrophic, ranging from operational disruptions and data breaches to financial losses and reputational damages. To mitigate these risks, companies are required to adopt a multi-layered cybersecurity strategy, which includes firewalls, Intrusion Detection Systems (IDS), and antivirus softwares. Moreover, cybersecurity frameworks and standards, such as National Institute of Standards and Technology (NIST) Cybersecurity Framework [11], provide guidelines for companies to systematically assess and manage their cybersecurity risks [12], [13]. Nevertheless, an alternative way to attack production plants consists in tampering hardware components by means of Hardware Trojan (HT) attacks [14], [15], [16]. Specifically, HT is a malicious modification enabled on physical devices, typically due to an intentional coupled circuit, that introduces a passive electronic component (such as a capacitor or a resistance) to disturb a voltage signal. HT attacks can also be introduced by connecting other circuits (such as other sensors) on the same electronic board, adding unnecessary electrical links on the same hardware platform or workstation, or intentionally altering the environmental conditions (as an example, a variation of temperature or humidity can drastically modify the electrical values of the circuit elements).

A preliminary study presented in [17] demonstrates that HT attacks can be modeled by modifying the output signal of a sensor by means of a HT resistance (RT) or capacitance (CT) added to the output of the analyzed circuit. Other forms of cyberattack can be modeled by adding a signal noise influencing electronic amplifier systems [18] or passive filtering networks [19]. These studies show that AI can provide important information to predict the signal distortion and hence adopt the corrective settings in machinery. In this context, the aim of this paper is to model a complex HT attack involving a sensor at the *level 1* of the Operational Technology (OT) layer of the Purdue Enterprise Reference Architecture (PERA) model [20]. In particular, as depicted in Fig. 1, the proposed approach combines the simultaneous effects of the HT resistance and capacitance, while also introducing a generic white noise signal controlled by a malicious user, which represents an advancement over the existing literature. The white noise is a random signal, commonly also named as background noise, presenting an even frequency distribution [18] that decreases the signal/noise ratio of circuits. White noise can be intentionally generated by electromagnetic interferences of other circuital elements positioned close to the sensor and generating inducted electric E and magnetic B fields. In addition, the presented model makes use of AI techniques to make predictions. Hence, by combining HT simulation and AI-based prediction in IIoT, this work offers a novel approach to securing critical industrial infrastructures against evolving hardware-level threats.

Summing up, the main contributions of this work are as follows:

- developing a unified AI-based digital twin model to simulate and predict HT attacks;
- considering possible malicious controlled white noises in the attacks;
- presenting a HT simulation environment to model attacks on IIoT hardware architectures;
- combining into the simulation model supervised learning AI techniques to predict and identify HT threats.

The effectiveness of the presented methodology is tested on a simulation environment to predict the output behavior of a sensor subjected to hardware level attacks and intentional noises and reply to such attacks. Nonetheless, we remark that, given the adopted black box approach, the proposed method can also be applied to reply to hardware trojan attacks to different circuits in any industrial application.

The remainder of the paper is structured as follows:

- section II summarizes the current state of the art for hardware security mechanisms and advanced AI techniques for HTs detection;
- section III introduces the circuital parametric analysis of the sensor modeling HT attacks and white noise malicious switching control;
- section IV describes the AI-based predictive models estimating the sensor output signals adding all the possible signal distortions (HT and white noise effects);
- section V reports discussions about the attack sensitivity and the execution of possible corrective actions compensating the HT attacks;
- section VI presents concluding remarks and future work directions.

II. RELATED WORK

The presented research aim at advancing the current state of the art on simulation and prediction of Hardware Trojan attacks in industrial applications by integrating AI-based prediction models with a comprehensive Hardware Trojan attack simulation framework tailored to the complexity of IIoT systems. In order to highlight the scientific advances of the proposed work with respect to the current literature on the topic, Table 1 presents a comparative analysis between the proposed work and other relevant contributions, including the early proposal described in [17]. For each selected publication, Table 1 reports the main findings, the adopted platform or device, the main focus of the proposed analysis, the presence or absence of noisy conditions in the study, and the adopted method for signal analysis and HT detection, also highlighting the main outcomes, limitations, and research gaps.

Initial research into HTs mainly focused on hardware-level detection methods, such as side-channel analysis and logic testing. These techniques exploit the abnormal behavior of HT-infected circuits, such as power consumption, timing variations, or electromagnetic emissions, to identify the presence of trojans [21]. While effective in controlled environments,

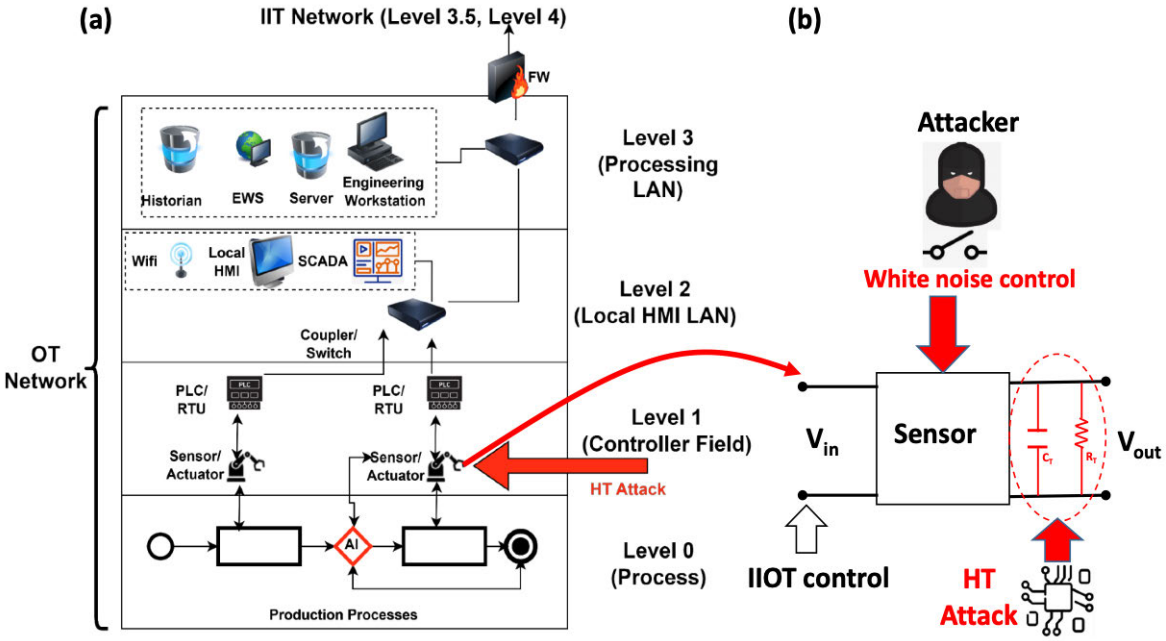

FIGURE 1. Level 1 of an industrial OT network where HT attacks are located (a) and sensor equivalent scheme (b) modelling HT attacks (RT and CT coupling) adding a controlled white noise signal disturbing the signal reading.

these methods often struggle to adapt to the complexity and variability of real-world IIoT systems, which are increasingly decentralized and heterogeneous [22]. Several frameworks for modeling and evaluating HT attacks have been proposed in prior works to facilitate a better understanding of the phenomenon and support the mitigation of these threats. For instance, a run-time monitoring technique to detect HTs in Field-Programmable Gate Arrays (FPGAs) was proposed in [23]. This work allowed researchers to study the impact of HTs on the performance of IoT devices. However, simulations were limited to specific hardware platforms thus being difficult to generalise to the diverse hardware architectures commonly adopted in IIoT systems. In addition, lacking of integration with AI-based prediction models, the proposed technique is unable to provide proactive defense mechanisms. A security framework was developed in [24] to detect HTs in different System-on-Chips (SoCs) platforms. This research advances IIoT security by offering an hardware-integrated approach to detect attacks and secure firmware updates. Nevertheless, the proposal neither use signal analysis for detecting HTs nor AI-based anomaly detection.

In this regard, the proposed work aims to address the limitations of traditional data analysis methods by integrating AI-based techniques. Indeed, AI and Machine Learning (ML) have shown significant promise in addressing security challenges in IIoT, including intrusion and anomaly detection [25], [26], particularly for identifying complex patterns within large datasets. Some recent works explored the integration of ML methods to identify cyber-physical attacks in industrial environments. In particular, the work in [27] demonstrated the efficacy of AI models, exploiting recurrent k-Nearest Neighbors (k-NN), Logistic Regression (LR), and Decision Tree (DT) classifiers to detect anomalies in IIoT networks. However, this approach, while effective for network-level attacks, do not directly address the specific challenges of hardware-level threats, such as HTs. Further approaches introduced Deep Learning (DL) to enhance HT detection. For instance, a DL-based framework was proposed in [28] to analyze power side channel data and to detect vulnerabilities in integrated circuits. The proposal significantly improved real-time detection accuracy with respect to other ML models, including Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs), but was limited to specific attack scenarios and its scalability to large-scale IIoT networks was not thoroughly investigated. By analyzing Table 1 we can conclude that previous works typically simulated scenarios under simple, controlled conditions without accounting for noisy environments, which limits their real-world applicability. In contrast, the proposed approach aims to provide a more comprehensive modeling and simulation framework that incorporates complex, and potentially malicious, noisy interferences. In this way, it enhances the accuracy of threats detection and response strategies in dynamic and less predictable settings. Summing up, unlike prior efforts, this research aims at developing a unified simulation and prediction framework combining: (i) a realistic Hardware Trojan simulation environment that can model attacks on IIoT hardware architectures; (ii) AI-driven prediction models designed to identify HT threats proactively, using supervised learning techniques.

TABLE 1. Comparative analysis and scientific advances of the proposed work (✗: not considered, potential area for further investigation).

Specification	Proposed work	Hemavathy et al. 2025 [24]	Massaro et al. 2024 [17]	Cheng et al. 2024 [23]	Saraf et al. 2024 [27]	Bhatta et al. 2024 [28]
Research finding	AI-based simulation and prediction framework for HT detection	security framework for detecting HTs in heterogeneous SoCs	methodology for modeling the effects of HT attacks on industrial control systems	run-time monitoring framework for HT detection on FPGAs used in IoT device	HT detection in integrated circuits in critical infrastructure scenarios	ML-based approach for HT identification using power side-channel signals
Reference platform/device	HR202 resistive humidity sensor module	Zynq UltraScale+ Multiprocessor SoC	HR202 resistive humidity sensor module	Xilinx SoC FPGA used in IoT sensor networks	Xilinx SoC FPGA	Xilinx SoC FPGA
Focus of parametric and sensitivity analysis	effect of simultaneous resistance and capacitance alterations in HT attacks	impact of HTs on resource utilization of FPGA components	impact of HT resistance and gain variation on PID controller performance	impact of HT attacks on the power consumption of a device	variations in the power consumption of a ring oscillator physical unclonable function	effect of different features (statistical, temporal, spectral) of side-channel data in HT attacks
Analysis under noisy conditions	attacker switching and modulation introducing a continuous white noise signal at the circuit input	different temperatures and voltages used to simulate real-world noisy conditions	✗	denoising process applied for outlier removal and mean filtering	✗	✗
Methods for signal analysis and HT detection	AI-based using ANN and RF algorithms implemented in KNIME	numeric metrics (utilization scale, digest uniqueness) used to identify functional alterations	✗	residual analysis with autocorrelation based on detection thresholds	AI-based using DT, K-NN and LR algorithms	AI-based using DT, RF, SVM, ANN algorithms
Main outcomes	AI algorithms accurately capture HT patterns, corrective actions suggested to compensate the impact of HT attacks	lightweight and efficient authentication protocol implemented for HT detection and firmware security	analog signals can be accurately modeled in OT networks for detecting HT attacks	statistical models show low hardware overhead with high detection accuracy	comparison of ML classifiers indicates high accuracy values for DT and LR models	identification of three HT states (enabled, disabled, triggered) correlated to key features of input signal
Limitations	further analysis required for managing potential bias into sensor measurements	lacks mechanisms to detect or prevent physical tampering attacks	noisy signals caused by unmanaged HT attacks, potentially leading to circuit defect	HT detection focused to specific platforms and based on simple evaluation of threshold values	simple binary HT classification with no prediction strategies	limited model generalizability due to a restricted range of nuanced differences in the dataset
Research gaps	distributed learning on edge devices, interpretability of results produced by AI algorithms	side-channel monitoring of signal behavior to identify changes in physical properties	automatic detection of HTs using AI algorithms	multi-modal detection techniques to improve adaptability of the proposal	analysis of the relationship between input attributes and model outcomes	dynamic adaptation to noisy scenarios, evaluation of model robustness in real-world use cases

III. CIRCUITAL PARAMETRIC ANALYSIS MODELING HARDWARE TROJAN SENSITIVITY

The proposed circuit model reported in Fig. 2 (a) is an extended version of the preliminary study performed in [17], based on a *HR202* resistive humidity sensor module (see Appendix A) for more details). The HT can be modeled by adding to the circuit two elements, that is, a HT resistance RT = R4 and a HT capacitance CT = C3, as shown on the right side of Fig. 2 (a).

It is to be noticed that parameters R4 and C3 are efficient in modeling the HT effects. Indeed, this part of the circuit is able to modify the output of the electrical signal providing voltage charge and discharge effects, that are respectively described by the following equations:

$$V_{out} = V_{input}(1 - e^{-t/(R_4C_3)}) \quad (1)$$

$$V_{out} = V_{input}(e^{-t/(R_4C_3)}) \quad (2)$$

The interested reader is referred to [17] for more details.

As illustrated in Fig. 2 (a), with respect to the previous work in [17] the sensor module now also includes a controlled white noise signal adding its effect on the simultaneous effects of the HT resistance and capacitance.

The switching control of the white noise signal follows the periodicity of the input pulsed signal that characterizes the IIoT triggering. Fig. 2 (b) and Fig. 2 (c) respectively illustrate the simultaneous effect of RT and CT with and without the noise switch control. The circuit simulation is performed on the LTSpice[1] open-source tool, calculating all the electrical variables of the analyzed circuit and executing the parametric simulations. The switch control is a pulsed signal that can be phase-shifted if compared with the input periodic signal. In the proposed analysis, the white noise signal is activated during the IIoT transmission window. When the noise signal's switching period matches the transmission window, it results in complete coverage of the reading intervals.

A preliminary parametric analysis is conducted to understand the effects of the HT attack, excluding the disturbances caused by white noise. Without HT attack, the voltage output of the *HR202* humidity sensor is reported in Fig. 3 (a). The simultaneous effect of CT and RT in the HT attack is shown in Fig. 3 (b) by changing the values of the electrical elements as follows: CT = C3 = 400 μF, 4 μF, 40 μF, and RT = R4 = 100 Ω, 1 $k\Omega$, 100 $k\Omega$, 400 $k\Omega$.

The voltage output of the *HR202* sensor, simulated by introducing a continuous white noise signal at the circuit input, is shown in Fig. 4. The white noise is commonly used to describe a random noisy process defined as a random process $X(t)$ having a flat power spectral density $S_X(f)$:

$$S_X(f) = constant \quad \text{for all frequencies} \tag{3}$$

Hence, for the purpose of our study, it is preferable to simulate chaotic hardware attack processes by introducing signal disturbances across the entire spectrum in which all possible circuits of controlled machines operate.

IV. THE AI-BASED PREDICTIVE MODEL

The predictive model is implemented by adopting the Konstanz Information Miner (KNIME) tool.[2] The ML algorithms used for the analysis are the Artificial Neural Network (ANN) [29] and the Random Forest (RF) [30], since they are commonly used for the prediction and classification of data achieved by circuital simulations, and they result particularly suitable for predicting noisy signals with low error rate, as demonstrated, for instance, in [18] and in [19]. As a matter of fact, ANN and RF algorithms exhibit performance processing simulation data that are very close to those of circuits influenced by noises [19].

The reference dataset, simulated using LTSpice, consists of 5,051,356 records. This extensive dataset captures a wide range of scenarios and conditions, allowing for a comprehensive evaluation of the system behavior under different parameters. The data input of the ANN and RF algorithms are the simulated voltage signals reported in Fig. 4, while the output is the predicted output voltage, selecting as target class a specific parametric circuit solution. The simulation data serves as a robust foundation for training and testing the AI models, ensuring that the predictive algorithms are exposed to diverse input variations and potential disturbances. According to Table 2, 8 different configurations have been defined to perform an early comparison of the exploited ML models. The ANN (see Appendix B for further details) is characterized by the following hyperparameters: 70% training dataset (linear sampling), 30% of the testing dataset (linear sampling), a maximum number of 400 iterations (epochs), 7 hidden layers with 25 neurons per layer. The RF network is characterized by the following hyperparameters: a tree depth of 10 levels, a minimum node size of 5, and 100 generated models. For each configuration, the ANN and RF results are obtained by fixing the HT parameters and using different values for C3 and C4. The results of the evaluation between the LTSpice simulated outputs of the circuit in Fig. 2 (a) without the application of the white noise signal and the corresponding values predicted using KNIME are respectively reported in Fig. 5 and in Fig. 6.

A good match is observed between the simulated results and the predicted values, with a low prediction error, indicating that both the ANN and RF algorithms are effective in modeling the system. The results highlight the ability of both algorithms to accurately capture the underlying patterns in the data, despite the presence of noise and potential disturbances.

Table 3 presents the comparison of the average error parameters between the ANN and RF algorithms. In particular, Table 3 reports the following metrics to preliminarily evaluate the model performance:

TABLE 2. Configuration for comparative analysis of ML models.

ID	ML algorithm	C3 (μF)	R4 ($k\Omega$)	Figure
1	ANN	40	1	Fig. 5 (a)
2	RF	40	1	Fig. 5 (b)
3	ANN	4	1	Fig. 5 (c)
4	RF	4	1	Fig. 5 (d)
5	ANN	400	100	Fig. 6 (a)
6	RF	400	100	Fig. 6 (b)
7	ANN	40	100	Fig. 6 (c)
8	RF	40	100	Fig. 6 (d)

TABLE 3. Error rates of ML models.

Error Parameter	ANN	RF
Mean Absolute Error (MAE)	0.003	0.013
Mean Squared Error (MSE)	0.001	0.001
Root Mean Squared Error (RMSE)	0.010	0.019
Mean Signed Difference (MSD)	0.002	0.005

[1] https://www.analog.com/en/resources/design-tools-and-calculators/ltspice-simulator.html

[2] https://www.knime.com

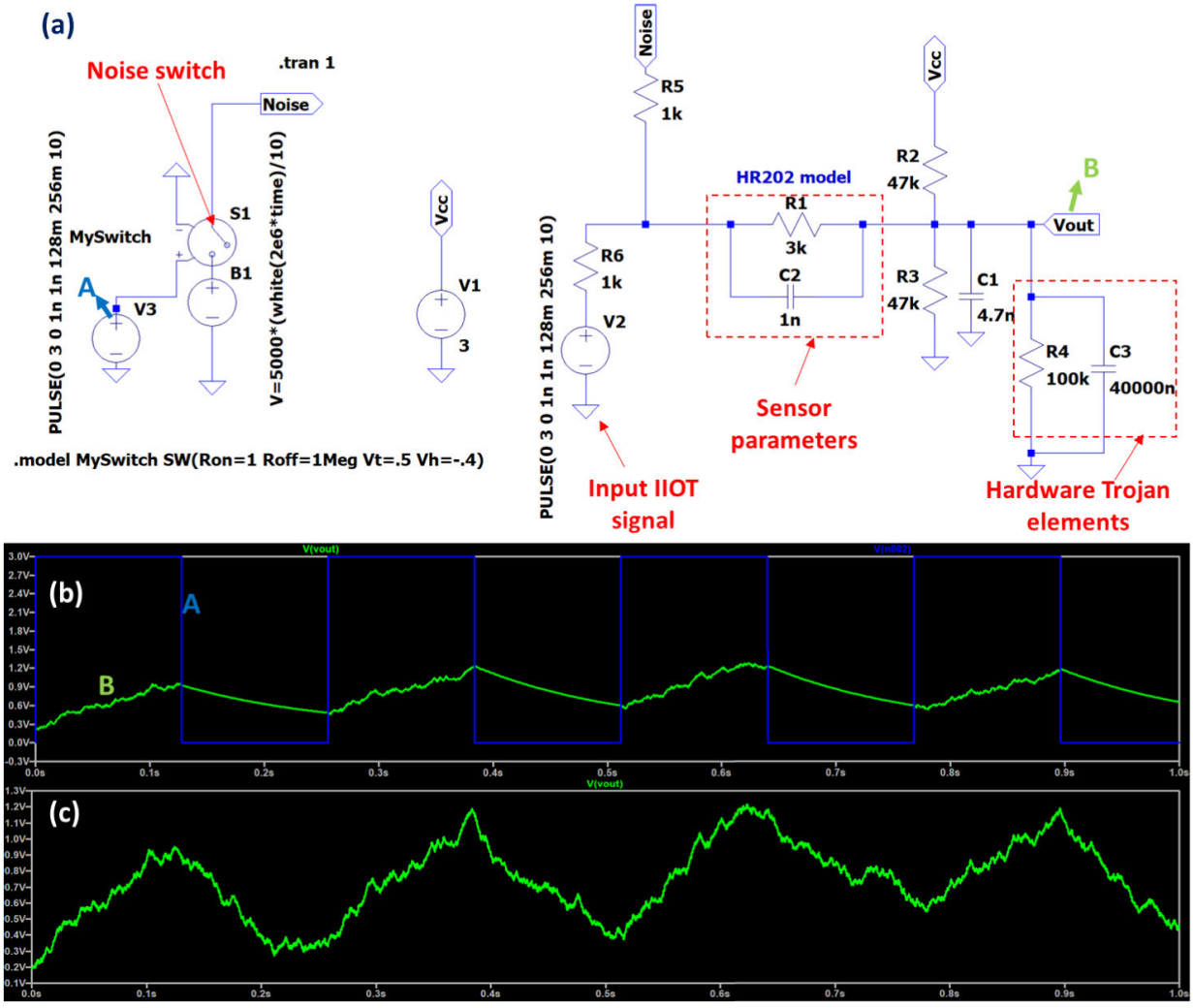

FIGURE 2. (a) Simulated sensor *HR202* circuit under HT attack (simultaneous effect of HT resistance and capacitance due to a hardware attack). (b) Switch signal (A: blue rectangular pulse) controlling HT, and voltage output affected by the white noise (B: green voltage output signal). (c) Voltage output signal without the switching signal control A (continuous effect of the white signal noise).

- Mean Absolute Error (MAE) to measure the average magnitude of errors in predictions, providing a clear understanding of the model's accuracy;
- Mean Squared Error (MSE) to quantify the average of the squared differences between predicted and actual values, providing a more sensitive measure for larger errors.;
- Root Mean Squared Error (RMSE) to penalize larger errors and evaluate the model's performance on a scale that is comparable to the magnitude of the data;
- Mean Signed Difference (MSD) to examine the average of signed differences between predictions and actual values, which helps understand whether the model tends to over or under-predict.

Each of these metrics serves a specific purpose in evaluating model accuracy, and their thresholds vary based on industry standards, sensor precision, and operational constraints. In particular, MAE and RMSE are often compared to sensor error tolerances to determine practical usability. MSE is used to fine-tune models, ensuring minimal large deviations, whereas MSD helps detect systematic prediction biases, ensuring reliability in long-term deployment. By benchmarking these metrics against industry-specific error tolerances and operational constraints, we ensure that the AI-based predictive models in IIoT systems meet practical performance expectations and can be confidently deployed in real-world applications. In the early evaluation, both algorithms exhibit strong performance, with minimal discrepancies between the predicted and actual values, which justifies their selection for the analysis. In particular, the ANN algorithm, with its ability to learn complex non-linear relationships, and the RF algorithm, known for its robustness and interpretability, both demonstrate that they are suitable for the given task.

The second step of the analysis is to predict the output voltage also by adding the white noise signal. We remark that we trained the supervised AI model through a parametric

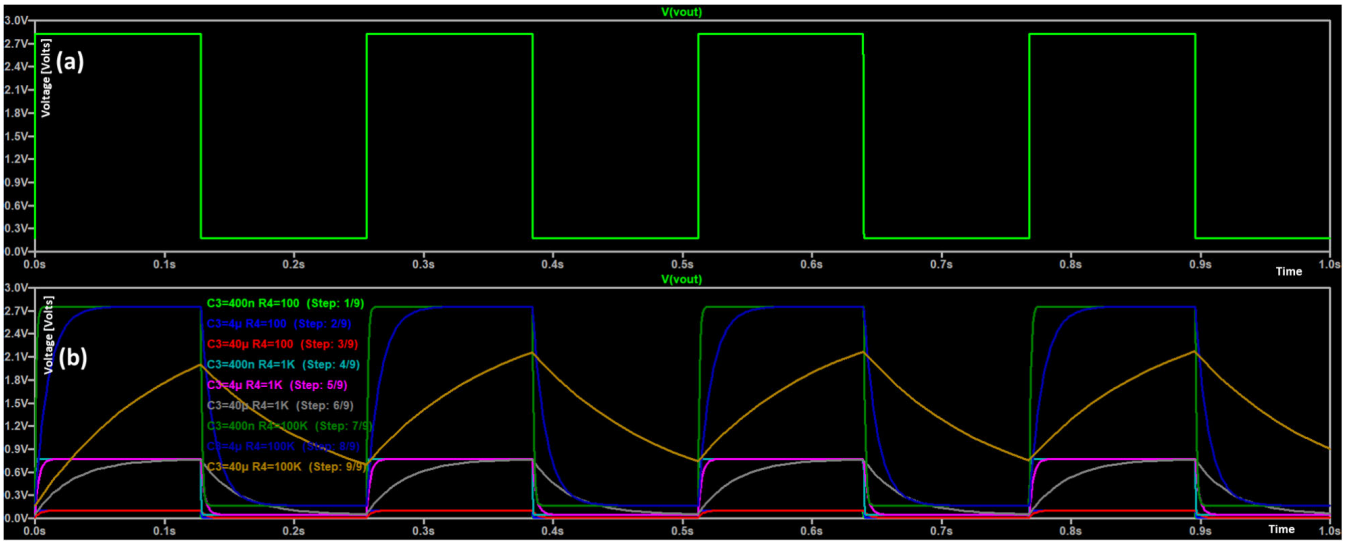

FIGURE 3. (a) Example of *HR202* voltage output without the white noise effect and the HT attacks. (b) Parametric analysis related the simultaneous effects of RT and CT without the superposition of the white noise.

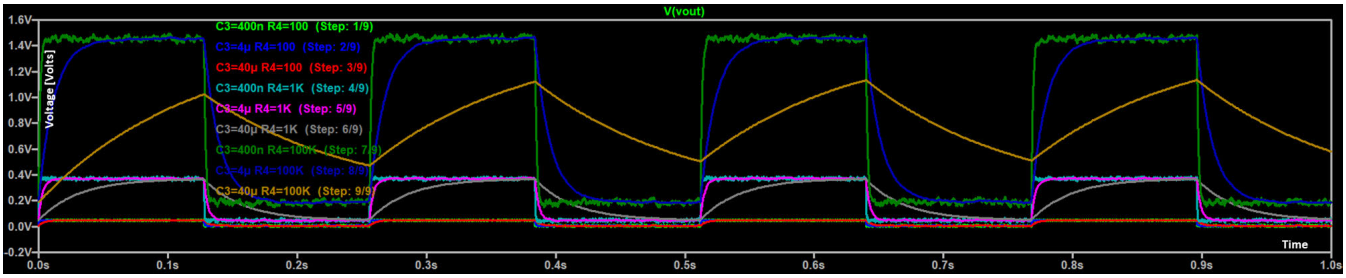

FIGURE 4. *HR202* sensor voltage output: voltage parametric analysis of the HT attack adding a continuous white signal.

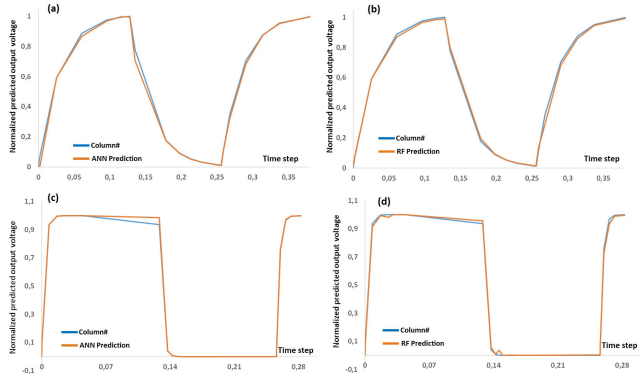

FIGURE 5. Comparison of predicted and simulated outputs for configurations 1-4 in Table 2.

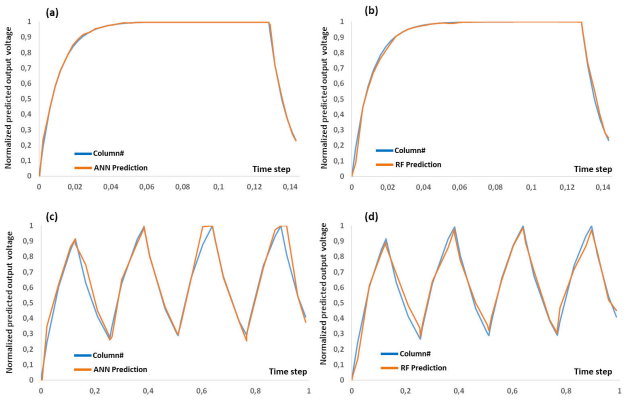

FIGURE 6. Comparison of predicted and simulated outputs for configurations 5-8 in Table 2.

analysis, assuming a range of circuit behavior based on noise amplitude modulation. By modifying the parametric analysis, it is easily possible to realize different training models that characterizes other possible sensor behaviors. For this reason, the LTSpice tool is here adopted to construct the specific AI training model to predict the output voltage. Focusing on the ANN algorithm, Fig. 7 shows the match between the simulated signal and the predicted signal for the ANN network trained starting from a basic configuration (see Fig. 7 (a)) and for the optimized one (see Fig. 7 (b)) using the same hyper-parameter configuration already described. As can be observed in Fig. 7 (b), the optimized algorithm is also able to follow the trend of ripple of the noisy profile, thus

facilitating the interpretation and analysis of the hardware attack. Furthermore, the matching between the simulated and the predicted results, along with the calculated error rates, confirms that the LTSpice simulation data are well-suited for training the AI predictive model. The simulated data represents a realistic use case, ensuring that the models are trained on a controlled yet practical dataset. The low prediction error observed in both algorithms suggests that the simulated data capture the key characteristics of the system, allowing the models to generalize effectively. This demonstrates that the LTSpice simulations can serve as a reliable foundation for building predictive models. However, it is worth mentioning that, in real-world applications, several factors can introduce bias into sensor measurements, affecting the accuracy and reliability of predictive models. Bias can stem from sensor calibration issues, data sampling limitations, environmental and contextual factors, and temporal variations. These biases may reduce model generalization and lead to misleading business decisions. To mitigate their impact, techniques such as sensor recalibration, balanced data sampling, feature normalization, periodic model retraining, and cost-sensitive learning should be applied. Addressing these challenges ensures robust and fair predictive modeling in real-world scenarios.

V. DISCUSSION

The effect of the HT resistance and capacitance is to change the amplitude profile of the output voltage of the *HR202* sensor. As can be observed in Fig. 3, the signal amplitude tends to decrease when the R4 value is decreased from 100 $k\Omega$ to 100 Ω, besides the simultaneous effect of the C3 value is to change the transient amplitude profile (the lower value of C3 $= 400\ \mu F$ corresponds to the best reading condition). Part of image in Fig. 3 is zoomed in Fig. 8, fixing R4 $= 100\,k\Omega$. The correct reading region is highlighted by the colored arrows while varying the C3 parameter. In particular, C3 $= 40\ \mu F$ corresponds to a full wrong reading condition. The same Fig. 8 also indicates the different sampling times of the LTSpice tool ranging from 40 μs to 4 ms.

By changing the value of R4, a sensitivity analysis of the voltage response can be also performed. Indeed, as reported in Fig. 4, the parametric analysis used to train the ANN model considers both a simple amplitude variation and cases of amplitude variation with a ripple behavior (that is, the white noise effects). By doing so, the ANN training model takes into account all the different signal profiles influencing the output, which in turns increases the model's robustness. Fig. 4 enhances that the white noise effect is more evident for specific values of C3 and R4. In particular, the noise amplitude varies with the C3 and R4 values and can be also modulated by the R5 resistance modeling the white noise intensity (as demonstrated by the ripple behavior in Fig. 2), which could be controlled by an attacker.

Fig. 9 highlights the modulation effect of R5, ranging from 1 Ω to 50 Ω. The fixed HT circuital parameters are R4 $= 100\ k\Omega$ and C3 $= 40\ \mu F$. It is to be noticed that,

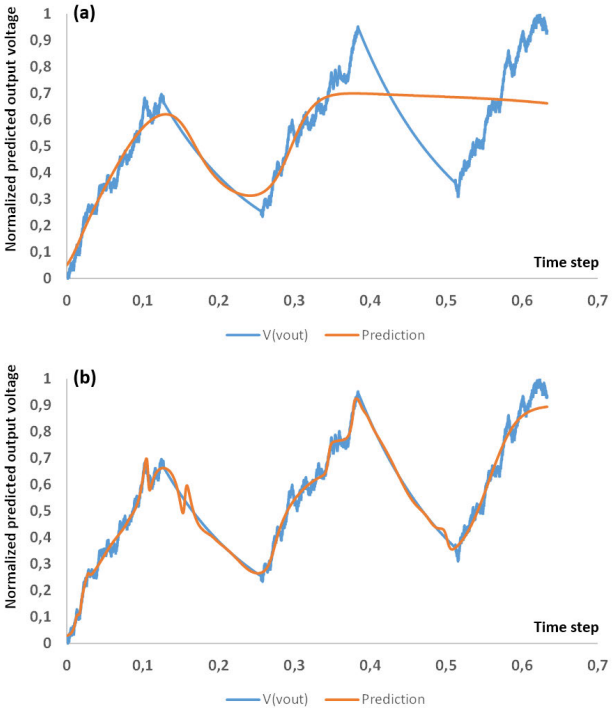

FIGURE 7. Comparison of predicted and simulated (named V(vout)) output voltage of the circuit sketched in Fig. 2 (a) with the application of a switch controlled white noise signal: (a) basic ANN network; (b) optimized ANN network (see section B).

by modifying the value of R5, it is possible to define different levels of signal degradation. A value of R5 $= 1\,\text{k}\Omega$ is close to real scenarios (that is, not accentuated and presenting ripples and not flat response). Nonetheless, the robustness of the proposed model also allows estimating the voltage outputs in case of accentuated ripples. The parametric analysis of the voltage estimated at the exit of the resistance R5 is shown in the Fig. 10, where R5 $= 1\,\text{k}\Omega$ is reported in the red plot).

Based on the simulated and predicted results, it is possible to identify specific corrective actions to compensate the attack effects. Specifically, the following automatic actions are assumed:

- *simple amplification*: when only a decrease in voltage amplitude is both detected and predicted (large correct reading region), the corrective action involves simple amplification, adjusting the gain to compensate for the decreased amplitude gap;
- *amplification and amplitude profile compensation*: when a signal with a substantial amplitude decrease and a change in the voltage output profile is detected and predicted, the corrective action involves simultaneously adding a circuit network to adjust the profile and an amplification stage;
- *multi-stage filtering*: in the presence of a rippled signal caused by a strong white noise effect, a multi-stage passive Resistance/Capacitance (RC) network can be applied to compensate for the ripple [19]. Additionally,

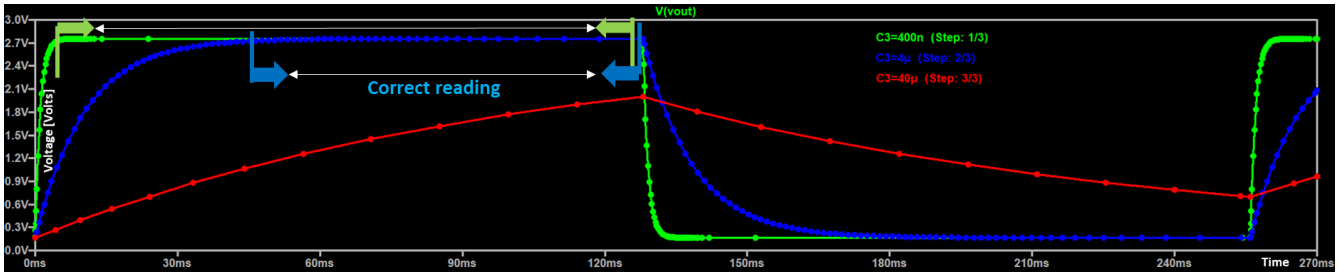

FIGURE 8. Zoomed plot of Fig. 3. The correct reading region is highlighted by the colored arrows.

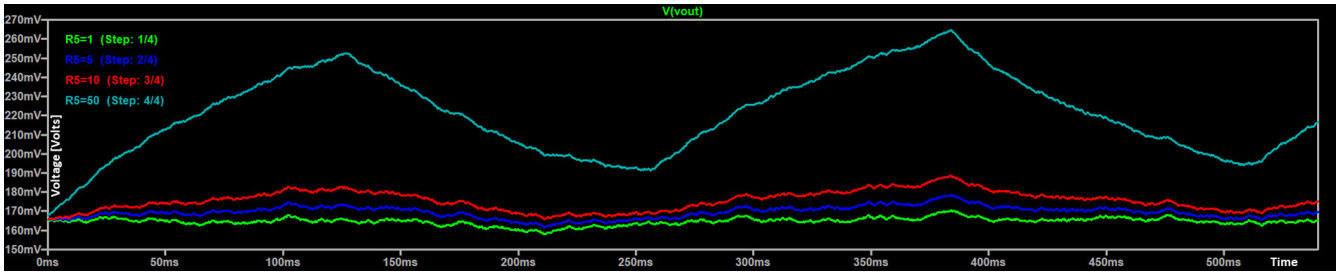

FIGURE 9. Modulation of the white noise effect by changing the resistance R5 of the circuit in Fig. 2(a).

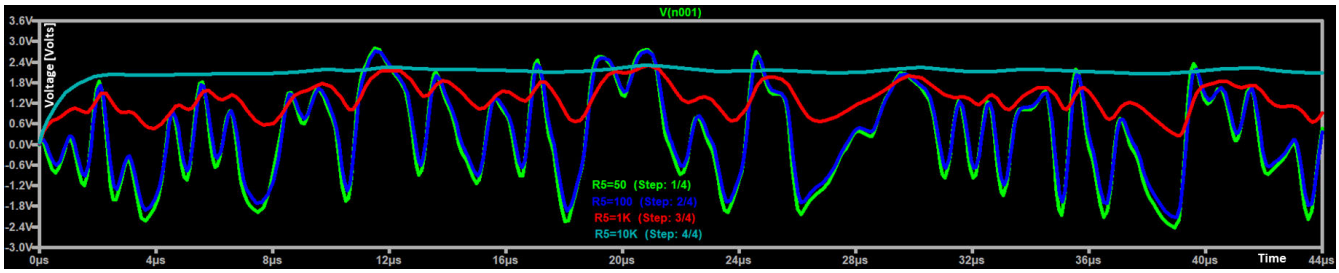

FIGURE 10. Parametric analysis of the voltage estimated at the exit of the resistance R5 for the circuit in Fig. 2 (a) (R5 = 1 kΩ is reported in red plot).

due to a possible further decrease in voltage amplitude, it may also be necessary to amplify the filtered signal;

- *signal shifting*: when a phase shift in the voltage output signal is detected, inductance and capacitance elements can be applied to correct the phase of the output signal. A phase shift in the output signal could be detrimental to the transmission of periodic pulsed signals, such as those used in IIoT systems;
- *waiting a response*: when the detected signal is regular and the predicted one is irregular, the corrective action is deferred until the observed voltage signal profile can be classified into one of the previously described cases.

When the above listed effects are not detected but predicted, it is suggested to anticipate the corrective actions as soon as anomalies are observed [19]. In this regard, a suitable approach may consists in gradually tuning the voltage signal based on the most recently observed values.

One of the main limitations of the proposed model is its difficulty in distinguishing between the effects of white noise and those of high-temperature interference, which often result in a decrease in the amplitude of the sensor output voltage. Consequently, the model operates with corrective actions based solely on the voltage output trend, aiming to compensate for the overall effect of the disturbance.

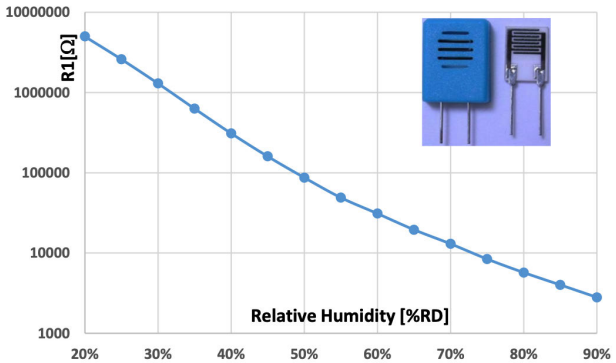

FIGURE 11. Characteristics of the *HR202* resistive sensor: R1 resistance of circuit Fig. 2 versus the relative humidity expressed in percentage. Inset: photo of the *HR202* sensor.

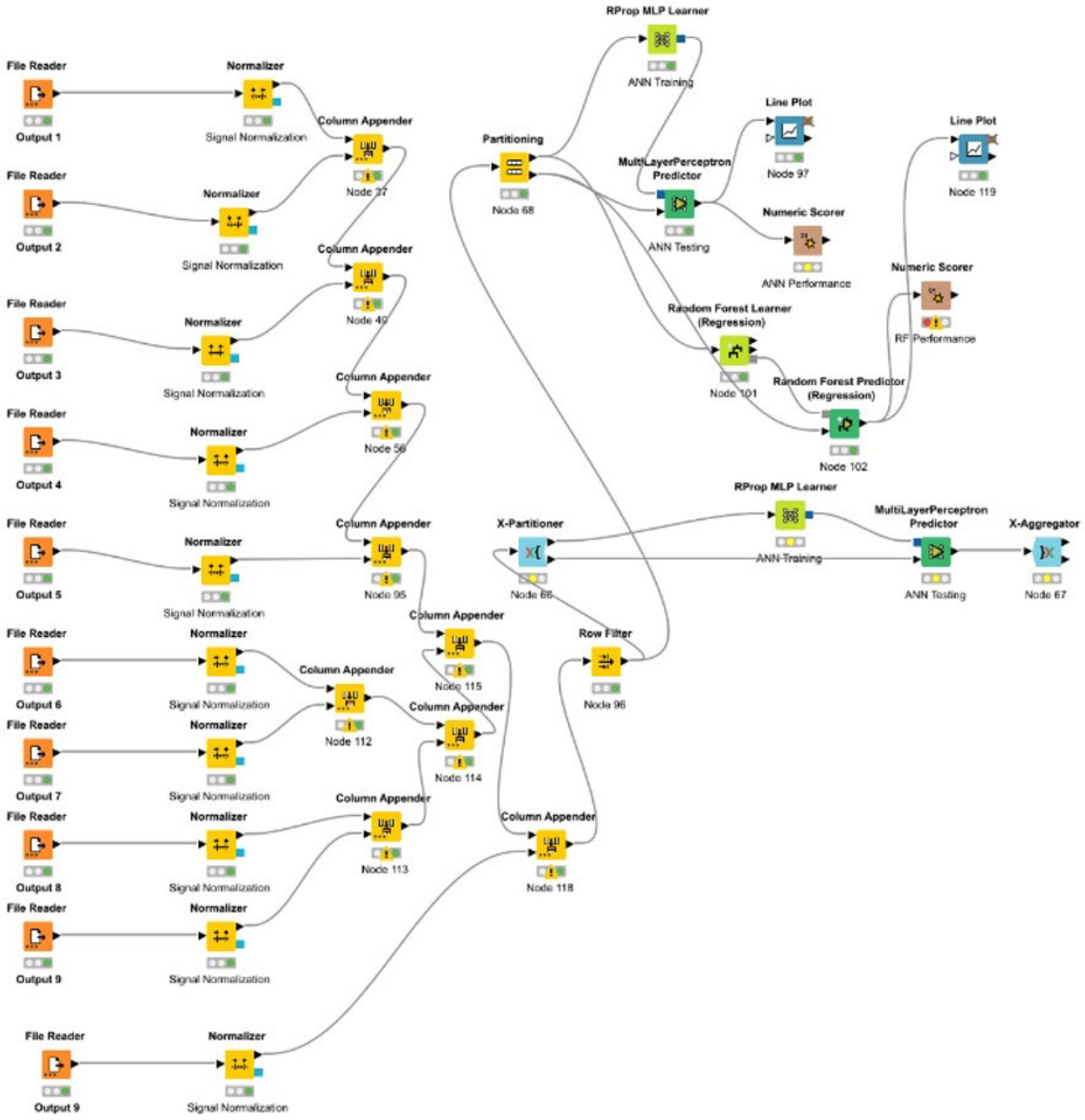

FIGURE 12. KNIME workflow implementing ANN and RF algorithms.

Furthermore, although the ANN models demonstrated strong performance in capturing complex relationships, its inherent lack of interpretability remains a key limitation. Unlike decision tree-based models such as RF, which provide clear feature importance rankings and decision paths, ANN functions as a *black box*, making it difficult to directly understand how inputs influence predictions. This lack of transparency can reduce trust in predictive maintenance models and anomaly detection systems, especially in mission-critical industrial environments where machine failures can lead to costly downtime or safety risks. If end-users cannot understand why a model detects a specific outcome or event, they may hesitate to act on its recommendations, limiting adoption. To overcome these limitations, eXplainable AI (XAI) techniques, such as Shapley Additive Explanations (SHAP), can be adopted to interpret model decisions and enhance trust and transparency in ML-based HT detection. XAI methods provide human-interpretable reasoning for flagged anomalies, improving cybersecurity response. In HT detection, interpretable insights can be used to identify which sensor readings or circuit parameters contribute most to detecting hardware anomalies or understand whether power

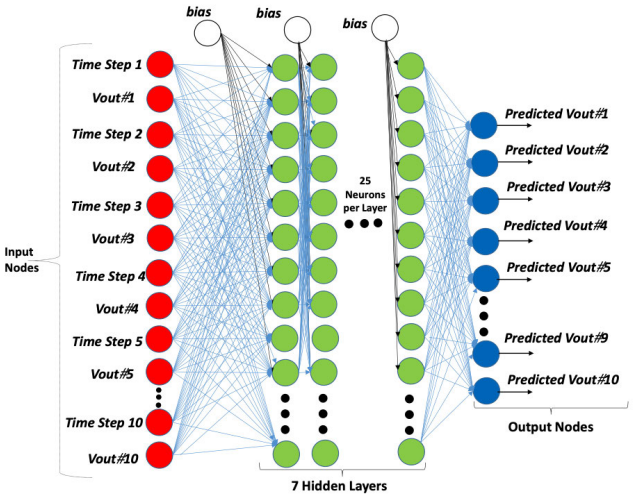

FIGURE 13. Architecture of the optimized ANN network.

consumption, timing delays, or signal variations indicate a cyber attack.

We conclude this section highlighting that, depending on the training model (that is, through the combination of different electrical parameters as performed in the parametric simulations), the proposed digital twin is able to predict not only the sensor output but also possible HT attacks with intentional noise interferences, which would otherwise be difficult to detect. The proposed model is adaptable to different parameter settings, allowing it to simulate and predict a wide range of hardware attacks.

VI. CONCLUSION

The paper discusses an innovative methodology to characterize different forms of hardware attacks in OT level 1 of industrial networks. As an example, the paper analyzes the sensitivity of the attacks varying HT electrical parameters and controlling a malicious white noise of a humidity sensor. Furthermore, the study proposes different corrective actions to compensate the impact of potential hardware attacks. High performance ANN and RF algorithms are trained to provide another possible way to collect information about the predicted behavior of the attack and to compensate preliminary the attack actions ensuring a good IIoT transmissions. By implementing the predictive model into the KNIME tool, a good match is observed between the simulated results (obtained through the LTSpice software) and the predicted values. Both the ANN and RF algorithms show their ability in accurately capturing the underlying patterns in the data, despite the presence of noise and potential disturbances. Results show the effectiveness of the proposed approach in predicting HT attacks, even in the presence of intentional interferences from the attackers

The proposed approach can be also applied to different sensors monitoring production variables in the whole working frequency range. The AI application moves the research topic from Industry 4.0 to Industry 5.0 scenarios by providing important elements to construct HT digital twin models. The development of a complete IIoT testbed platform based on the coupling of micro-controllers and sensors is currently under investigation.

A further promising direction for future research is the integration of federated learning approaches for HT attack simulation and detection. In traditional centralized AI models, data from different IIoT devices must be aggregated in a central location, raising privacy and scalability concerns. Federated Learning (FL) offers a decentralized alternative by enabling individual IIoT devices to collaboratively train AI models without sharing their local raw data [31]. However, while FL enhances privacy of data, model parameters exchanged during training can still be vulnerable to inference attacks. Ensuring robust encryption and secure aggregation is crucial for protecting sensitive industrial data. FL can also improve the generalization of AI models in resource-constrained IoT environments, where real-time threat detection is critical, as they would learn from heterogeneous hardware environments and diverse attack patterns. Edge devices in IIoT can have limited processing capabilities and memory, making FL training computationally expensive. In these scenarios, lightweight model architectures, such as pruned neural networks and quantized models, are suggested to reduce computational demands. Implementing asynchronous FL updates can further optimize communication efficiency by reducing synchronization delays also avoiding network bandwidth congestion. Future work will also focus on using Reinforcement Learning (RL) to model evolving HT behaviors [32]. Dynamic attack scenarios will be simulated in which HTs change their activation conditions or evade detection over time. This approach will test the performance of the RL-based detection system to assess whether it can generalize beyond known attack patterns by introducing previously unseen HT with novel activation mechanisms. In addition to basic metrics related to prediction accuracy, the following evaluation criteria will be used to quantify effectiveness of RL models: (a) adaptation speed, i.e., the number of training episodes required for the RL agent to adjust to a newly introduced attack strategy; (b) robustness to concept drift, to determine the model's ability to maintain performance when the underlying data distribution shifts; (c) sensitivity to noise, for assessing how well the model can handle noisy or incomplete data, which is common in real-world applications. This could provide deeper insights into how future trojans might adapt and evade identification procedures, leading to the development of more proactive and intelligent HT defenses in IIoT systems. Finally, we plan to extend the contribution by comparing the accuracy of the proposed prediction approach with that of other prediction methods.

APPENDIX A

HR202 is a resistive sensor detecting the relative humidity of the air. The basic model of this sensor is a circuit composed by a resistance (see Fig. 2 (a)) in parallel with a capacitance

TABLE 4. Error rates of the cross-validation test (10 folds).

Fold Number	Total Squared Error	Mean Squared Error
0	0.025	0.002
1	0.034	0.003
2	0.016	0.001
3	0.037	0.003
4	0.039	0.003
5	0.032	0.003
6	0.033	0.003
7	0.037	0.003
8	0.034	0.003
9	0.039	0.003

C2 ranging between 0.43 μF to 7.6 μF that characterizes the charging time. The electrical resistance R1 varies with the detected relative humidity according to the characteristic that can be observed in Fig. 11: the sensor output voltage V_{out} is a function of the electrical resistance R1 and, consequently, of the relative humidity [17].

APPENDIX B

The KNIME workflow used for AI-based prediction is illustrated in Fig. 12. The input nodes consist of the voltage outputs from the LTSpice simulations discussed in section III. The workflow implements both the ANN and RF algorithms. Furthermore, the k-fold cross-validation algorithm [33] (with k = 10) is applied to evaluate the performance of the ANN. Following to the RPROP (resilient propagation) algorithm, the loss function used to train the ANN model is the error function that quantifies the difference between the predicted output and the actual target value [34]. The architecture of the optimized ANN is shown in Fig. 13. The performance results of the cross-validation are summarized in Table 4, demonstrating the efficiency of the linear sampling.

REFERENCES

[1] A. Massaro, *Electronics in Advanced Research Industries: Industry 4.0 to Industry 5.0 Advances*. Hoboken, NJ, USA: Wiley, 2021.
[2] M. Ragnoli, M. Pavone, N. Epicoco, G. Pola, E. De Santis, G. Barile, and V. Stornelli, "A condition and fault prevention monitoring system for industrial computer numerical control machinery," *IEEE Access*, vol. 12, pp. 20919–20930, 2024.
[3] M. Dotoli, N. Epicoco, M. Falagario, and G. Cavone, "A timed Petri nets model for intermodal freight transport terminals," *IFAC Proc. Volumes*, vol. 47, no. 2, pp. 176–181, 2014.
[4] G. Cavone, M. Dotoli, N. Epicoco, M. Franceschelli, and C. Seatzu, "Hybrid Petri nets to re-design low-automated production processes: The case study of a Sardinian bakery," *IFAC-PapersOnLine*, vol. 51, no. 7, pp. 265–270, 2018.
[5] S. Ieva, D. Loconte, G. Loseto, M. Ruta, F. Scioscia, D. Marche, and M. Notarnicola, "A retrieval-augmented generation approach for data-driven energy infrastructure digital twins," *Smart Cities*, vol. 7, no. 6, pp. 3095–3120, Oct. 2024.
[6] S. Ashraf and T. Ahmed, "Sagacious intrusion detection strategy in sensor network," in *Proc. Int. Conf. U.K.-China Emerg. Technol. (UCET)*, Aug. 2020, pp. 1–4.
[7] M. Dotoli and N. Epicoco, "Integrated network design of agile resource-efficient supply chains under uncertainty," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 11, pp. 4530–4544, Nov. 2020.
[8] S. Akter, K. Khalil, and M. Bayoumi, "A survey on hardware security: Current trends and challenges," *IEEE Access*, vol. 11, pp. 77543–77565, 2023.
[9] M. Lezzi, M. Lazoi, and A. Corallo, "Cybersecurity for industry 4.0 in the current literature: A reference framework," *Comput. Ind.*, vol. 103, pp. 97–110, Dec. 2018.
[10] F. Akbarian, W. Tärneberg, E. Fitzgerald, and M. Kihl, "Attack resilient cloud-based control systems for industry 4.0," *IEEE Access*, vol. 11, pp. 27865–27882, 2023.
[11] D. P. Möller, "NIST cybersecurity framework and MITRE cybersecurity criteria," in *Guide to Cybersecurity in Digital Transformation: Trends, Methods, Technologies, Applications and Best Practices*. Cham, Switzerland: Springer, 2023, pp. 231–271.
[12] J. Li, J.-J. Qiu, Y. Zhou, S. Wen, K.-Q. Dou, and Q. Li, "Study on the reference architecture and assessment framework of industrial internet platform," *IEEE Access*, vol. 8, pp. 164950–164971, 2020.
[13] K. Stouffer, K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, and A. Sherule, "Guide to operational technology (OT) security," Dept. U.S. Commerce, Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. 800-82r3, 2023.
[14] M. M. R. Monjur, J. Heacock, J. Calzadillas, M. S. Mahmud, J. Roth, K. Mankodiya, E. Sazonov, and Q. Yu, "Hardware security in sensor and its networks," *Frontiers Sensors*, vol. 3, May 2022, Art. no. 850056.
[15] M. R. Monjur, S. Sunkavilli, and Q. Yu, "ADobf: Obfuscated detection method against analog trojans on I2C master-slave interface," in *Proc. IEEE 63rd Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2020, pp. 1064–1067.
[16] M. Monjur, J. Calzadillas, and Q. Yu, "Hardware security risks and threat analyses in advanced manufacturing industry," *ACM Trans. Design Autom. Electron. Syst.*, vol. 28, no. 5, pp. 1–22, Sep. 2023.
[17] A. Massaro, N. Epicoco, G. Loseto, and C. Ardito, "IIOT cyberattacks in industrial field control systems: PID attack and hardware trojan effects," in *Proc. IEEE 20th Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2024, pp. 1702–1707.
[18] A. Massaro, "ANNs predicting noisy signals in electronic circuits: A model predicting the signal trend in amplification systems," *AI*, vol. 5, no. 2, pp. 533–549, Apr. 2024.
[19] A. Massaro, "Artificial intelligence enabling denoising in passive electronic filtering circuits for industry 5.0 machines," *Machines*, vol. 12, no. 8, p. 551, Aug. 2024.
[20] T. J. Williams, "The Purdue enterprise reference architecture," *IFAC Proc. Volumes*, vol. 26, no. 2, pp. 559–564, Jul. 1993.
[21] H. Li, Q. Liu, and J. Zhang, "A survey of hardware trojan threat and defense," *Integration*, vol. 55, pp. 426–437, Sep. 2016.
[22] S. Sidhu, B. J. Mohd, and T. Hayajneh, "Hardware security in IoT devices with emphasis on hardware trojans," *J. Sensor Actuator Netw.*, vol. 8, no. 3, p. 42, Aug. 2019.
[23] J. Cheng, Q. Feng, C. Li, and W. Yang, "Securing FPGAs in IoT: A new run-time monitoring technique against hardware trojan," *Wireless Netw.*, vol. 30, no. 6, pp. 5455–5463, Aug. 2024.
[24] S. Hemavathy, K. Jagadeesh, and V. S. K. Bhaaskaran, "Unified security framework using device-specific fingerprint: Mitigating hardware trojans and authenticating firmware updates," *IEEE Access*, vol. 13, pp. 26897–26914, 2025.
[25] Z. Huang, Q. Wang, Y. Chen, and X. Jiang, "A survey on machine learning against hardware trojan attacks: Recent advances and challenges," *IEEE Access*, vol. 8, pp. 10796–10826, 2020.
[26] K. I. Gubbi, B. Saber Latibari, A. Srikanth, T. Sheaves, S. A. Beheshti-Shirazi, S. M. Pd, S. Rafatirad, A. Sasan, H. Homayoun, and S. Salehi, "Hardware trojan detection using machine learning: A tutorial," *ACM Trans. Embedded Comput. Syst.*, vol. 22, no. 3, pp. 1–26, May 2023.
[27] M. Saraf, T. H. Syed, A. Kulkarni, and M. Niamat, "Hardware trojan detection employing machine learning, physical unclonable functions and side channel analysis," in *Proc. IEEE Int. Conf. Electro Inf. Technol. (eIT)*, May 2024, pp. 514–519.
[28] N. P. Bhatta, U. Giri, and F. Amsaad, "Machine learning-based classification of hardware trojans using power side-channel signals," in *Proc. IEEE 67th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2024, pp. 990–994.
[29] J. Zou, Y. Han, and S.-S. So, "Overview of artificial neural networks," in *Artificial Neural Networks: Methods and Applications*. Totowa, NJ, USA: Humana Press, 2009, pp. 14–22.

[30] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble Machine Learning: Methods and Applications*. New York, NY, USA: Springer, 2012, pp. 157–175.
[31] D. Loconte, S. Ieva, A. Pinto, G. Loseto, F. Scioscia, and M. Ruta, "Expanding the cloud-to-edge continuum to the IoT in serverless federated learning," *Future Gener. Comput. Syst.*, vol. 155, pp. 447–462, Jun. 2024.
[32] Z. He, H. Homayoun, and H. Sayadi, "Beyond conventional defenses: Proactive and adversarial-resilient hardware malware detection using deep reinforcement learning," in *Proc. 61st ACM/IEEE Design Autom. Conf.*, Jun. 2024, pp. 1–6.
[33] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Statist. Comput.*, vol. 21, no. 2, pp. 137–146, Apr. 2011.
[34] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *Proc. IEEE Int. Conf. Neural Netw.*, Jun. 1993, pp. 586–591.

ALESSANDRO MASSARO (Senior Member, IEEE) carried out scientific research with the Polytechnic University of Marche, CNR, and Italian Institute of Technology (IIT) as the Team Leader by activating laboratories for nanocomposite sensors for industrial robotics. He is also register in MIUR as a Scientific Expert in competitive industrial research and social development. He was the Head of the Research and Development Section and the Scientific Director of the MIUR Research Institute, Dyrecta Laboratory Srl. He is also a member of the International Scientific Committee of Measurers IMEKO and IEEE TC34 Technical Committee Secretary, received an award from the National Council of Engineers as a Best Engineer of Italy, in 2018. He was a Research and Development Manager of LUM Enterprise, in 2022 and 2023. Actually, he is an Associate Professor in electronics with the Department of Engineering, LUM University of Casamassima, Italy. He has ING-INF/01, FIS/01, and FIS/03.

NICOLA EPICOCO (Member, IEEE) received the master's degree in mechanical engineering and the Ph.D. degree in electrical and information engineering from the Polytechnic of Bari, Italy, in 2002 and 2016, respectively. From 2019 to 2022, he was Assistant Professor with the University of L'Aquila, Italy, and in 2021, he was a Visiting Researcher with the École Centrale de Lille, France. He is currently an Associate Professor in control systems engineering with LUM University, Bari, Italy. He has authored more than 50 international publications. His research interests include modeling, simulation, optimization, and control of dynamic systems. He is involved in the following technical committees: Intelligent Systems to Human-aware Sustainability of the IEEE SMC Society, Digital Manufacturing and Human-Centered Automation of the IEEE RAS, and Transportation Systems (TC 7.4) and Control for Smart Cities (TC 9.3) of the IFAC. He is also involved in international and national research projects and in the program committee of several international conferences.

GIUSEPPE LOSETO (Member, IEEE) received the master's and Ph.D. degrees in information technology engineering from the Polytechnic University of Bari, in 2009 and 2013, respectively. He is currently an Associate Professor with the Department of Engineering, LUM University. His research interests include pervasive computing and the Internet of Things, knowledge representation systems, and applications for ubiquitous smart environments. On these topics, he has co-authored over 70 papers in international journals and conferences. He is involved in international and national research projects and is a program committee member of several international conferences and workshops. He is also an editorial board member of journals in areas related to his research interests.

CARMELO A. ARDITO received the master's degree in computer science and the Ph.D. degree in computer science from the University of Bari "Aldo Moro." He is currently a Full Professor with the Department of Electrical and Information Engineering, Polytechnic University of Bari. His research interests include human–computer interaction, with a particular focus on design, development and evaluation of new interaction techniques and pervasive systems (multimodal interaction with mobile systems, interaction through gestures with large multitouch screens and mixed reality, and interaction through "tangibles" with smart environments), human-centred artificial intelligence, and cybersecurity. He is the co-author of more than 130 papers published in journals, international conferences, and collections. He is an Expert Member of the International Federation for Information Processing Technical Committee on Human–Computer Interaction (IFIP TC13), the Deputy-Chair of the IFIP TC13 Working Group "13.2—Methodologies for User-Centered Systems Design," a member of the IFIP IoT Domain Committee, and a member of the ACM SIGHCI and of the Italian Chapter on Human–Computer Interaction.