Société d'Automatique,
de Génie Industriel & de Productique

Enhancincing Data Integrity: A Solution to Predict Data Disturbance in IoT Systems

Meriem Smati[1,2], Vincent Cheutet[1], Jannik Laval[1], Christophe Danjou[2]

[1] INSA Lyon, Université Lumière Lyon 2, Université Claude Bernard Lyon 1, Université Jean Monnet Saint-Etienne, DISP UR4570, Villeurbanne, 69621, France

[2] Laboratoire Poly-Industrie 4.0, Département de Mathématiques et Génie Industriel, Polytechnique Montréal, Montréal, Québec, Canada

meriem.smati@insa-lyon.fr, vincent.cheutet@insa-lyon.fr, jannik.laval@univ-lyon2.fr, christophe.danjou@polymtl.ca

Keywords : *Digital Twins, Internet of Things, Model-Driven Engineering, Machine Learning, Data Integrity.*

1 Introduction

The integration of cognitive capabilities into Digital Twins (DTs) has led to the emergence of Cognitive Digital Twins (CDTs)[4], enhancing their intelligence and predictive abilities. A CDT refers to a virtual representation of a physical or digital entity that possesses the ability to perceive, analyze, learn, and make decisions in a manner similar to human cognition, it leverages Artificial Intelligence (AI) and Machine Learning (ML) to replicate the cognitive processes and behaviors of its real-world counterpart. This study explores its potential in strengthening the resilience and maintenance of Internet of Things (IoT) systems [1]. Moreover, a new framework, Cognitive Super Digital Twin (CSDT), is being introduced that not only replicates the actions of the system but also generates data to simulate normal and abnormal scenarios. By combining cognitive technologies and augmented data generation, this study highlights the transformative role of CDTs in enhancing IoT system security amidst the increasing integration of IoT devices. So instead of relying solely on traditional methods such as redundancy [2], the focus is put on leveraging digital technologies like CDTs to provide a cost-effective solution and utilise AI and ML to mimic human cognition to ensure system's reliability and informed decision-making. The research objective is to answer the following research question:

How can DT be used as a tool for detecting and preventing data disturbance in an IoT system?

2 CSDT Proposal

As shown in Figure 1, DTs combine a set of layers (illustrated in pink). The database layer stores and manages data by using two repositories "Dataset Acquisition Repository" and "Vault Repository", while the simulation layer uses the data to simulate and replicate the behavior of the system. The visualization layer allows users to see a virtual and real-time representation of the system, and the decision layer helps humans make decisions. To make predictions and automate decision-making, a cognitive layer is added (represented in Orange), transforming the DT into a CDT.

But IoT systems are susceptible to a range of vulnerabilities stemming from both software and hardware failures considered as obstacles and challenges that encompass the uneven distribution of data in IoT systems, with a prevalence of normal data that complicates anomaly detection; the variability in defining normal and abnormal behavior, which is affected by factors like season, location, or context; and the dependence on historical data in traditional DT, requiring a data anomaly to occur before making predictions in future cases. This led to the addition of the Data Generation layer, colored in Blue, which makes the CDT a CSDT. The justification for this generation module is the lack of abnormal data compared to normal ones. This layer fabricates normal and abnormal data, both of those generations are essential to have a balanced global dataset. Therefore, distinguishing those two types of behavior is already a challenge [3]. For instance, a temperature of 10°C is considered normal during winter, but the same temperature in summer might be regarded as a deviation.

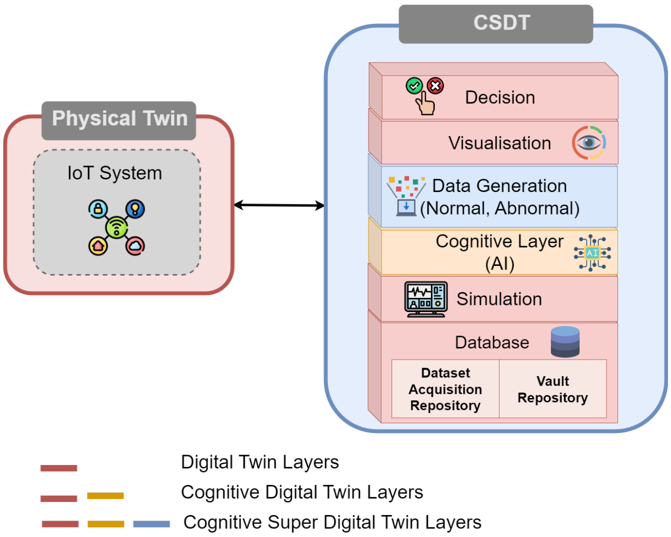

FIG. 1: CSDT Layers

3 Case Study

The contribution has been implemented in a case study to test the capabilities of CSDT that takes into consideration potential disturbances such as irregularities in the data, which can significantly impact system performance, Assess the scalability of its simulator, identify suitable ML or DL models for both the Physical Twin (PT) and the CSDT. To mitigate the mentioned disturbances, data mining techniques like Classification-based anomaly detection can be employed, leveraging the presence of a data generation module. Disturbances primarily involve data manipulation, deletion, or simulated sensor halts, depending on the environment and sensor behavior. The use case is about detecting room occupancy by collecting environment data in a continuous, time-series stream. Contextual anomalies are present, and while the PT tackles a binary classification problem (Occupied, not Occupied), the CSDT addresses multiclass classification (Occupied, not Occupied, disturbed).

In the use case, the PT comprises two subsystems. The first subsystem adopts an Edge/Fog IoT architecture, featuring a Raspberry Pi 3 Model B and three sensors: Grove DHT11, Grove Light, and Grove SGP30. Later, a Grove Barometer sensor is added to demonstrate the extensibility of the CSDT's simulator. Data collected (temperature, humidity, CO2, and light) are sent to the Fog, along with an external dataset, to enrich the PT's training data. The model's predicted output and collected data are transmitted to the second subsystem, consisting of a Poppy Ergo Jr robot assisting in decision-making based on the received data. Simultaneously, data are stored in InfluxDB via the Communication Medium. Upon receiving all data, the Communication Medium forwards it to the DT, where the data acquisition module preprocesses and combines three data sources (DT-generated normal and abnormal data, external dataset, collected data) into a unified dataset. The simulator mimics the first subsystem's behavior, facilitating data transmission from edge to fog. A multi-class model is trained on the global dataset, periodically replacing the existing model in the PT.

To accomplish prediction, various ML and DL models like Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Naive Bayes, Recurrent Neural Network (RNN), and MultiLayer Perceptron (MLP) have been trained. The PT model with the highest performance was an RNN with Long Short-Term Memory Network (LSTM) layers. This model was chosen for its consideration of timestamps and performance. However, tree-based models outperformed the RNN in the context of CSDT, achieving 99% scores compared to the RNN's 67.03% F1 Score. The decision between accurate predictions and considering timestamps led to prioritizing accurate predictions in this study. Additionally, as noted in a related study, tree-based models excel over other models when working with tabular data.

4 Conclusion

This paper introduces the CSDT framework, aiming to enhance the preventative capabilities of DT. While DTs offer benefits, they require significant investments and should complement traditional maintenance practices. The CSDT extends beyond replicating the PT, detecting and generating perturbations to improve prevention and resilience strategies. However, limitations include constraints related to data distribution and environmental context awareness. Additionally, the CSDT's simulator can only handle numerical data from sensors, lacking support for non-numeric data like images or videos.

Future endeavors include transitioning towards System of Systems (SoS) structures, expanding the CSDT from a component twin to a system or process twin. This evolution involves transitioning from static to dynamic DTs that operate in real-time, integrating real-time training mechanisms for continuous improvement. And potential use cases include incorporating reinforcement learning for improved performance over time and analyzing sensor patterns to predict anomalies. Furthermore, testing different combinations of anomaly injections using Model-Driven Engineering (MDE) can enhance anomaly detection accuracy by considering various system factors comprehensively.

References

[1] Pavlos Eirinakis, Stavros Lounis, Stathis Plitsos, George Arampatzis, Kostas Kalaboukas, Klemen Kenda, Jinzhi Lu, Jože M. Rožanec, and Nenad Stojanovic. Cognitive digital twins for resilience in production: A conceptual framework. *Information*, 13(1), 2022.

[2] Aron Laszka, Waseem Abbas, Yevgeniy Vorobeychik, and Xenofon Koutsoukos. Integrating redundancy, diversity, and hardening to improve security of industrial internet of things. *Cyber-Physical Systems*, 6:1–32, 06 2019.

[3] Amitabh Mishra, Achraf Cohen, Thomas Reichherzer, and Norman Wilde. Detection of data anomalies at the edge of pervasive iot systems. *Computing*, 103(8):1657–1675, 2021.

[4] Xiaochen Zheng, Jinzhi Lu, and Dimitris Kiritsis. The emergence of cognitive digital twin: vision, challenges and opportunities. *International Journal of Production Research*, 60(24):7610 – 7632, 2022.