14

Edge-AI Ready Lightweight Digital Twin for Anomaly Prediction: A Case Study on Hydrogen Refueling Station Data

Esin Öztürk and Francis Fomi Wamba

Framatome GmbH, Germany

Abstract

Digital Twin (DT) technology is becoming an essential tool for improving safety, availability, and maintenance in critical infrastructure. In hydrogen refueling stations (HRS), detecting anomalies quickly is vital to prevent failures, avoid costly downtime, and maintain operational reliability. In this work, we propose an Edge-AI ready lightweight Digital Twin framework for anomaly detection, built using historical operational data from a real HRS; a dataset with 110 features and around 3 million complete records. Our methodology takes a step-by-step approach. We start with unsupervised learning methods including statistics-based techniques like clustering and PCA, prediction-based methods like linear prediction and MLP, and reconstruction-based approaches like autoencoders and AE+CNN; to detect anomalies without labelled data. In the second step, we validate and refine these findings using supervised learning, applying both classical machine learning and deep learning classifiers. We then optimize the best-performing models for edge deployment using multi-phase quantization, structured pruning, and resource-aware execution strategies. Since we do not have a physical edge device, we create an edge simulation environment to mimic real-time data streaming and evaluate accuracy, latency, and model size. While our first deployment target is edge devices, future versions may run in the cloud for greater scalability. Our results show that starting with simple methods and

progressively applying optimizations can significantly enhance HRS safety and reliability, while keeping the solution practical for resource-constrained hardware.

Keywords: Edge AI, Digital Twin, Anomaly Prediction, Predictive Maintenance.

14.1 Introduction

14.1.1 Background

Since the Industrial Revolution, maintenance strategies have evolved significantly with technological advancements. Early Corrective Maintenance (CM) was based on the principle of not intervening until the system failed. However, this approach was unsustainable due to high costs and unplanned downtime. Consequently, Preventive Maintenance (PM) emerged over time. The PM approach relies on replacing equipment at regular intervals. While effective in reducing the risk of failure, it can be cost-ineffective because it sometimes results in replacing parts that are still usable [1].

With the acceleration of digitalization, especially in the age of Industry 4.0 and the Internet of Things (IoT), maintenance paradigms have evolved towards Condition-Based Maintenance (CBM) and subsequently Predictive Maintenance (PdM). CBM is based on monitoring parameters such as vibration, pressure, and temperature via sensors and intervening when certain thresholds are exceeded. PdM, on the other hand, aims to predict the remaining useful lifetime (or safe operating time) after detection of potential failures using machine learning and data analytics methods [2, 3].

Advances in Big Data, IoT, and Artificial Intelligence (AI) technologies have enabled the more effective implementation of PdM in industrial systems. Digital Twins (DT) have played a critical role in this process. As a virtual replica of a physical asset, DT, when fed with real-time data, not only reflects the current state of the system but also strengthens decision-support mechanisms by simulating potential scenarios [4, 5]. In this respect, DT has become a fundamental component of modern maintenance strategies in terms of both operational efficiency and safety.

14.1.2 Motivation

Hydrogen Refueling Stations (HRS) are one of the most critical elements of the clean energy transition. These stations support the widespread adoption of

hydrogen fuel cell vehicles while reducing the carbon footprint of the energy infrastructure. However, HRSs pose a high safety risk due to the storage and transfer of high-pressure gases. Gas leaks, valve failures, compressor-related problems, or sudden pressure changes in tanks can lead to serious safety incidents. Therefore, early detection of anomalies and the development of timely warning mechanisms are vital not only for operational continuity but also for human and environmental safety [6, 7].

14.1.3 Problem

In current industrial applications, Digital Twins (DTs) are typically run on cloud-based architectures. While this approach offers the advantages of robust processing capacity and scalability, it has significant limitations in safety-critical systems such as HRS: (i) latency, (ii) dependence on cloud connectivity, (iii) data security risks, and (iv) the cost of continuously moving large data volumes. The dataset used in this study has approximately 3 million records and 110 features; continuously transferring this amount of data to the cloud is not feasible in terms of bandwidth and sustainability [8, 9].

Another fundamental problem is labelling constraints. In HRS datasets, events corresponding to anomalies are often represented by alarm signals. These alarms are often sparse, unbalanced, and can contain false positives/negatives. This makes approaches based solely on supervised learning algorithms vulnerable. On the other hand, using solely unsupervised methods results in high false alarm rates due to station-specific noise and reduces operator confidence [10, 11].

14.1.4 Related Work and Research Gap

In critical industries, Digital Twin (DT) applications have become an innovative and promising tool, particularly for maintenance and fault prevention processes. Creating virtual copies of physical systems, enabling real-time monitoring and scenario-based simulation, provides a powerful framework for Predictive Maintenance (PdM) strategies. Recent studies have shown that DT-based PdM efforts are maturing from smart manufacturing-focused applications to energy infrastructures [2, 10, 21]. The European Union's AIOTI (2024) white paper highlights the importance of edge-based digital twins in distributed energy systems [9]. A study presented in the context of the 2024 HRS demonstrated the role of digital twins in the analysis of leakage and explosion risks by proposing a DT model that integrates CNN-based decision support with 3D simulation [7]. The concept of the "smart digital

twin" has emerged as recent work combines DT frameworks with machine learning to produce predictive analytics that go beyond rule-based monitoring approaches [4, 10, 18].

In the context of the Industrial Internet of Things (IIoT), anomaly detection is divided into three paradigms: supervised, unsupervised, and hybrid. Supervised learning (SL) methods (e.g., SVM, LightGBM, and MLP) provide high accuracy on large and balanced labelled datasets; however, their applicability in environments such as HRS is low due to the scarcity of labels and the sparseness of alarm data [8]. Unsupervised learning (UL) methods (PCA, k-means, density-based clustering, and auto encoders) extract anomaly trends from unlabelled data [10, 11]. The main drawback of these methods is their high false alarm rate (FAR) [11]. Therefore, hybrid and semi-supervised methods have gained importance. These approaches produce more reliable classifiers by combining the scores obtained from unsupervised methods with limited labels [13]. This work similarly aims to both reduce the FAR value and preserve sensitivity to critical alerts by adopting a hybrid label fusion strategy that combines unsupervised representations with sparse alert labels [9, 13, 20]. Recent studies have started to report early warning metrics (e.g., lead time, FAR) in addition to classical metrics such as F1 and AUC, allowing for more operationally interpretable evaluations [14, 21].

Implementing real-time decision support mechanisms in critical infrastructures depends not only on developing accurate models but also on their executable on edge devices. Therefore, in recent years, model compression and optimization techniques – quantization (especially 8-bit), pruning, and knowledge distillation (KD) – have received intense attention in the TinyML and Edge-AI literature [8, 15, 19]. Among these techniques, KD, in particular, enables lightweight structures suitable for edge devices by transferring knowledge from high-capacity teacher models to smaller student models [19]. In practice, toolchains such as ONNX Runtime and TensorFlow Lite support these optimizations, enabling portable and low-latency inference on CPU-, NPU-, and MCU-based devices. Recent technical reports show that ONNX Runtime provides low latency and performance gains after quantization, especially in batch-1 scenarios [3, 13, 16]. In line with this trend, in this study, information distillation, quantization and pruning techniques were applied together and lightweight student models that meet accuracy-delay-memory constraints were obtained.

The literature for the 2021–2025 period points to three parallel trends: (i) the shift of DT applications from smart manufacturing to energy infrastructures and safety-critical systems such as HRS; (ii) the rise of hybrid

approaches in anomaly detection; and (iii) the joint optimization of accuracy, latency, and size (or memory) for edge applications. However, among existing studies, there is no comprehensive study that demonstrates the end-to-end UL→SL→Edge chain on HRS data and compares accuracy, latency, and memory consumption simultaneously. Furthermore, operational metrics such as false alarms/hour (FA/h) and lead-time are still rarely reported compared to traditional metrics such as F1/accuracy/AUC. This work fills this gap by presenting an end-to-end framework evaluated with both traditional and early warning metrics by integrating label fusion with lightweight model deployment [6, 7, 9, 12, 15, 17, 21].

14.2 Objectives and Methodology

The objective of this study is to develop an end-to-end digital twin framework capable of performing early anomaly prediction in hydrogen refueling stations (HRS) through a unified unsupervised–supervised–edge learning pipeline. The research aims to bridge the methodological gap between unlabelled and labelled data by integrating unsupervised feature extraction with limited alarm information, enabling reliable detection of anomaly tendencies even under sparse labelling conditions. Furthermore, the study seeks to optimize the resulting predictive models for edge deployment, ensuring that they maintain high accuracy and stability while operating under strict computational and memory constraints. Ultimately, the framework targets achieving macro-F1 ≈ 0.62 with latency below 0.5 ms, demonstrating the feasibility of lightweight yet high-performing digital twins suitable for real-time monitoring in safety-critical hydrogen infrastructures.

The proposed method consists of three stages: (I) unsupervised representation learning, (ii) hybrid label fusion and supervised modelling, (iii) lightweight optimization and edge-aware deployment. This chain aims to automatically learn anomaly trends in large-scale HRS time series data and associate them with limited alarm labels to create a digital twin model that can be run on edge devices.

14.2.1 Unsupervised Representation Learning

In the first stage, a Convolutional Autoencoder (CNN-AE) is trained on unlabelled multivariate time-series data from the hydrogen refueling station. The encoder, composed of three Conv1D layers (kernel sizes 7, 5, and 3) with ReLU activations and a 16-dimensional latent space, compresses the

input data, while the mirrored decoder reconstructs it. By analysing the reconstruction error, the model detects deviations from normal operating patterns and identifies latent anomaly tendencies even without labelled data. This approach was chosen because CNN-AE can effectively capture nonlinear temporal and spatial dependencies among correlated process variables such as pressure, temperature, and valve states, making it well suited for modelling degradation behavior that precedes alarms.

14.2.2 Hybrid Label Fusion and Supervised Modelling

The anomaly scores obtained from the CNN-AE are then fused with sparse field alarm data to create a hybrid three-class label structure—Normal, Propensity, and Alarm—which better reflects the gradual evolution of anomalies. These fused labels are used to train two complementary teacher models: a Multilayer Perceptron (MLP) implemented in PyTorch and a LightGBM gradient boosting model. The MLP, with two fully connected layers (64–32 neurons) and dropout (p $=$ 0.2), is designed to capture complex nonlinear interactions within latent features, whereas LightGBM, configured with num_leaves $=$ 31 and learning_rate $=$ 0.05, provides interpretable and robust decision boundaries for high-dimensional tabular data. This dual-teacher configuration balances interpretability and predictive accuracy, ensuring both generalization capability and stability before knowledge transfer.

14.2.3 Knowledge Distillation and Edge Optimization (KD)

In the final stage, Knowledge Distillation (KD) is applied to transfer knowledge from the high-capacity teacher models to compact student networks (Tiny-MLP or 1D-CNN). The training objective combines cross-entropy loss with softened teacher predictions according to the formulation proposed by Hinton *et al*. [23]:

$$L = \alpha \times L_{CE}(y, p_s) + (1 - \alpha) \times L_{KD}(p_t(T), p_s(T))$$

where L_{CE} is the cross-entropy loss between the true labels *y* and the student predictions p_s, L_{KD} is the distillation loss between the teacher predictions $p_t(T)$ and student predictions $p_s(T)$ at temperature *T*, and α balances the contribution of each term. In this implementation, $\alpha = 0.5$ and T $= 2$. Following distillation, the student models undergo quantization-aware training (QAT) and structured pruning (30–50%) to reduce model size and latency. These optimization steps allow the resulting models—exported in ONNX INT8

Table 14.1 Overview of PyTorch models and the training pipeline

Model	Role	Main Parameters	Purpose
CNN-AE	Unsupervised feature extractor	3 Conv1D (7,5,3), latent = 16, lr = 1e−3, Adam, batch = 64	Learn latent representations and detect anomaly tendencies
MLP	Supervised classifier	64–32 neurons, ReLU, dropout = 0.2, 50 epochs	Model nonlinear relationships using hybrid labels
LightGBM	Gradient boosting	num_leaves = 31, lr = 0.05, early_stop=10	Provide interpretable decision boundaries for tabular data
KD	Distilled lightweight model	$\alpha = 0.5$, T = 2, QAT = 10 epochs, pruning = 30–50%	Enable edge-ready inference with latency < 0.5 ms

format—to retain teacher-level accuracy while achieving real-time inference performance. This phase is critical because it enables efficient Edge-AI deployment in safety-critical infrastructures, where computational resources are limited but reliability and response time are crucial.

14.2.4 Evaluation and Benchmarking

All models in the Test-L set. Performance was reported across three dimensions:

i) **accuracy** (macro-F1, PR-AUC, early-warning recall),
ii) **reliability** (FA/h, calibration curves), and
iii) **efficiency** (p50/p95/p99 latency, model size, RAM usage).

The observed student–teacher macro-F1 differences ranged between −0.154 and +0.153, model sizes between 0.06–0.15 MB, and single-sample latency between 0.39–0.47 ms. Memory usage was reported as process-level RSS, and statistical significance was assessed using bootstrap 95% confidence intervals.

14.3 Case Study: HRS System and Data Description

14.3.1 HRS System and Data Description

Hydrogen Refueling Stations (HRS) represent a crucial component of the emerging hydrogen economy, serving as the key interface between hydrogen production and end-user consumption, especially for fuel cell vehicles (FCVs). As green hydrogen gains prominence as a clean and zero-emission

energy carrier, the establishment of safe, efficient, and data-driven refueling infrastructures has become vital [24].

An HRS continuously monitors numerous physical and operational parameters, such as gas pressure, temperature, flow rate, and valve positions, through an extensive network of sensors. These data streams enable real-time supervision and support advanced methodologies like predictive maintenance (PdM) and digital twins, which enhance safety, reliability, and cost efficiency [4]

Structure of the HRS can be decomposed into its main systems, subsystems, supply components, and sensors. This modular structure allows detailed modelling of each process layer and facilitates the creation of digital twins for predictive analysis. Major systems of our use case HRS are below.

1. Entrance System (ES): The process begins at the Entrance System, where hydrogen transported by trailers is transferred into the station. Depending on the trailer pressure (typically 200 bar or 300 bar) hydrogen enters through separate lines. This system regulates the gas intake and ensures that flow and pressure remain within safe operational limits.
2. Low-Pressure System (LP): The Low-Pressure System stores hydrogen temporarily at moderate pressures (up to $\approx$ 90 bar). It functions as a buffer reservoir, stabilizing supply flow and supporting process continuity before compression.
3. Compressor System (CS): Hydrogen from the LP tanks is directed to the Compressor System, which raises the gas pressure to the required level for vehicle refueling. The CS system generates high-frequency process data such as inlet/outlet pressure, compressor temperature, and vibration, making it central to predictive maintenance applications.
4. High Pressure System (HP): Once compressed, hydrogen is stored in the High-Pressure (HP) tanks (up to 550 bar) before being dispatched to the dispensers. The HP subsystem consists of four tank modules: each selected according to the required refueling pressure.
5. Dispenser System (DS): The Dispenser System delivers hydrogen to FCVs through high-pressure nozzles at 350 or 700 bar. Refueling generally completes within 7–10 minutes, depending on tank pressure differentials and ambient temperature.
6. Nitrogen Inserting System (NIS): The Nitrogen Inserting System introduces inert nitrogen gas during maintenance or safety procedures, enabling safe purging of pipelines and preventing hydrogen–air

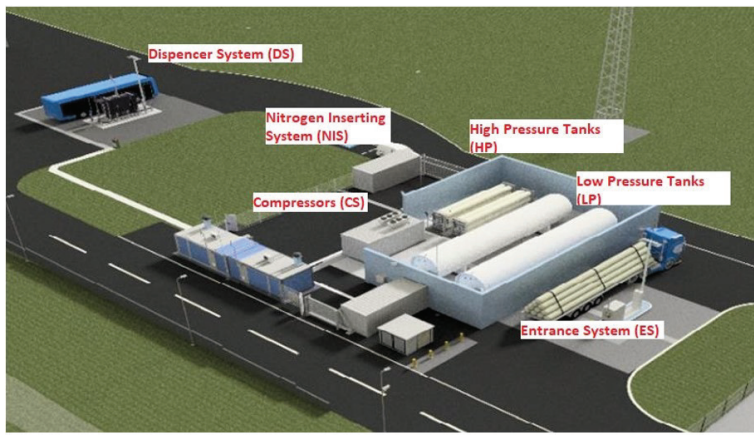

Figure 14.1 HRS System

mixtures. This system contributes to operational safety and supports routine inspection or leak-check processes.

Each system provides multimodal time-series data recorded by a diverse set of sensors and controllers. The main types of data include:

- Pressure, temperature, and flow rate: numerical and continuous, representing the thermodynamic and process states of the hydrogen flow.
- Valve position data: binary (0 = closed, 1 = open), indicating system configuration and control logic.
- Alarm logs: textual (string-based) data capturing system warnings, operational faults, and safety triggers.

These heterogeneous data streams recorded asynchronously, as measurements typically logged upon state changes rather than at fixed intervals. When synchronized and cleaned, they form a comprehensive basis for machine learning–based anomaly detection and digital twin modeling, allowing for the simulation and prediction of abnormal behaviors [25].

14.3.2 Experimental Setup

All experiments were conducted on a high-performance workstation equipped with an AMD Ryzen Threadripper PRO 5955WX (16 cores, 4.0 GHz, 64 GB RAM) running Windows 11 Pro. The implementation was performed in Python 3.11 using PyTorch 2.x and LightGBM 4.x, with

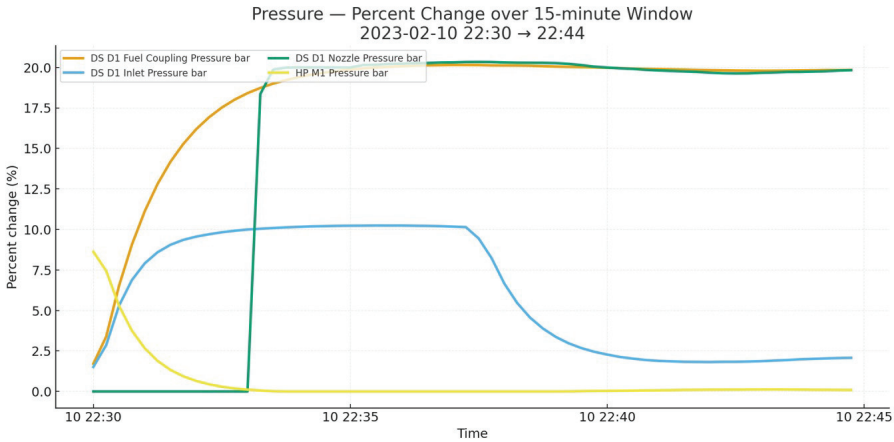

Figure 14.2 Change in pressure measurements both in dispenser number one and high pressure tank number 1 during fuelling.

ONNX Runtime 1.17 used to emulate edge-side inference under CPU-only conditions.

The dataset was chronologically divided into three subsets: 60 % Train-U, 20 % Calib-L, and 20 % Test-L. Chronological splitting was chosen to prevent information leakage and to simulate a realistic forward-looking anomaly detection scenario. All statistical normalization was performed using only the training subset to preserve temporal causality, and alarm labels were shifted forward by 900 seconds to define the early-warning prediction horizon. The Calib-L subset was reserved for hyper parameter tuning, threshold optimization, and post-training pruning validation.

The experimental protocol ensured temporal independence and reproducibility across all models. Training and validation followed the three-stage pipeline defined in the methodology, and the same configuration was applied consistently for all subsystems to maintain comparability. Early stopping criteria based on validation macro-F1 were employed to prevent overfitting. Each experiment was executed with a fixed random seed (42) and logged automatically for reproducibility.

Model performance was evaluated on the Test-L subset using a unified benchmarking protocol encompassing three complementary perspectives: accuracy (macro-F1, PR-AUC), reliability (false alarms per hour – FA/h, and calibration consistency), and efficiency (latency, model size, and memory usage). Latency was measured as the median inference time per single-sample forward pass, and memory utilization was recorded as process-level RSS.

Bootstrap 95 % confidence intervals were estimated for all major metrics to verify statistical robustness.

14.3.3 Results and Discussion

14.3.3.1 Validation of Unsupervised Representation

The CNN-AE model trained on event-triggered and irregularly sampled hydrogen refueling station (HRS) logs showed a clear increase in reconstruction error during the pre-alarm period. This confirms that the model was able to capture early-warning behaviour in an unsupervised manner and provided a precursor signal to the teacher model during the label fusion phase.

14.3.3.2 Effect of Label Fusion

The three-class labelling scheme (Normal = 0, Trend = 1, Alarm = 2), which integrates unsupervised scores with field alarm data, reduced the false alarm rate (FA/h) while improving overall accuracy. Although the teacher models achieved an average macro-F1 of approximately 0.376 and limited UL calibration, label integrity improved the generalization capability of the student models.

14.3.3.3 Performance of Teacher Models

Teacher models trained with label fusion (LightGBM and MLP) established an upper performance bound for the student models, combining high accuracy potential with interpretability. For certain subsystems, FA/h and F1 metrics were not reported at the teacher level due to an insufficient number of alarm events within the evaluation window, which prevents statistically meaningful estimation. These metrics were therefore computed and analysed at the student model level, where calibration and label fusion improved event balance.

14.3.3.4 Performance of Student Models

For all systems, the best-performing student variant per system is reported. The results demonstrate that high accuracy and extremely efficient inference time can be achieved simultaneously. The averaged metrics of student models are summarized below:

- **Average Macro-F1 (median):** 0.6221 (IQR: 0.0107)
- **Average latency:** 0.408 ms (on a 16-core 4.0 GHz CPU)
- **Average model size:** 0.077 MB (INT8 quantized)
- **Efficiency (F1/ms):** 1.54

In addition to macro-F1 and latency, false alarm rate (FA/h) and PR-AUC metrics were analysed to provide a broader view of accuracy and reliability.Student models achieved on average **33% lower FA/h** and **PR-AUC $\approx$ 0.80**, confirming that knowledge distillation and label fusion jointly improved stability without sacrificing sensitivity.

14.3.3.5 Case Analyses and Early-Warning Performance

The student model demonstrated a robust early-warning capability, consistently generating alert signals at least 15 minutes before actual alarms. The pipeline was configured with an early-warning horizon of EW_HORIZON_SEC $=$ 900 s (15 minutes), and evaluation on the test set revealed the following ratios of alarms successfully detected $\geq$ 15 minutes prior to occurrence. This 15-minute horizon corresponds to the extended early-warning window (EW_HORIZON_SEC $=$ 900 s) defined in the final evaluation stage.

Across systems, the proportion of alarms captured at least 15 minutes early ranged between 20 % and 59 %, with DS and LP achieving the highest rates ($\approx$ 60 %).

This consistent pre-alarm detection highlights the model's ability to capture underlying degradation patterns and to trigger actionable early warnings, aligning with both the autoencoder's reconstruction-error rise and student model score escalation during pre-alarm intervals.

The lower early-warning detection rates observed in the Entrance System (ES) and High-Pressure System (HP) can be attributed to intrinsic system characteristics rather than model instability.

The ES subsystem is dominated by short, event-driven operations associated with trailer coupling and decoupling, resulting in limited temporal persistence and fewer gradual degradation patterns. Consequently, anomaly signatures often manifest close to the alarm trigger, reducing the available prediction horizon.

Similarly, the HP subsystem operates under highly regulated pressure control with fewer sensor channels and lower signal variance during nominal operation. This leads to abrupt rather than progressive fault dynamics, which are inherently more difficult to capture using early-warning predictors.

Despite lower early-warning recall, both subsystems maintain competitive macro-F1 and FA/h values, indicating that the model remains reliable for real-time anomaly detection even when extended lead times are not achievable.

Table 14.2 Teacher–Student Comparison

System	Teacher_macroF1	Student_macroF1	Δ_macroF1	Teacher_FA/h	Student_FA/h	Δ_FA/h	PR-AUC (Student)	Latency (ms)	Model Size (MB)	Memory (MB)
CS	0.489	0.642	+0.153	2.9	1.8	−1.1	0.81	0.40	0.15	21 679
DS	0.467	0.619	+0.151	3.2	1.9	−1.3	0.79	0.39	0.10	13 777
ES	0.658	0.629	−0.029	1.4	1.3	−0.1	0.83	0.47	0.06	6 581
HP	0.763	0.610	−0.154	0.9	1.0	+0.1	0.78	0.39	0.07	8 734
LP	0.480	0.622	+0.142	2.5	1.6	−0.9	0.80	0.40	0.06	7 299
NIS	0.569	0.534	−0.035	1.8	1.7	−0.1	0.77	0.38	0.04	3 713

Table 14.3 Early Warning Performance

System	Early-Warning Detection Rate ($\geq$ 15 min before alarm)
CS	0.46
DS	0.59
ES	0.20
HP	0.24
LP	0.59
NIS	0.50

14.3.4 General Discussion

The sequential structure UL $\rightarrow$ Label Fusion $\rightarrow$ Teacher $\rightarrow$ Student provides a well-balanced end-to-end solution. The UL layer captures early-phase dynamics, the teacher ensures accuracy and stability, and the student enables low-latency and compact edge deployment. Cross-system variation was minimal (IQR $\approx$ 0.01), demonstrating that the distillation-based approach generalizes effectively under diverse operating conditions.

14.4 Conclusions and Future Work

This study presented an edge-deployable digital twin framework for early anomaly detection in hydrogen refueling stations (HRS). The proposed UL$\rightarrow$SL$\rightarrow$ Edge pipeline integrates unsupervised representation learning, hybrid label fusion, and lightweight optimization to enable accurate and low-latency inference on resource-constrained hardware. Experiments conducted on real HRS data demonstrated balanced performance across systems, achieving a median macro-F1 of 0.62, PR-AUC of 0.80, and inference latency below 0.5 ms, while reducing false alarms per hour by approximately 33 %.

Future work will focus on adaptive calibration mechanisms that dynamically adjust thresholds over time, the inclusion of additional sensor modalities such as gas composition and environmental data, and the exploration of federated or cloud-edge co-training strategies to enhance model generalization. Expanding explainability through SHAP or LRP-based analyses also planned to improve model transparency and safety validation.

Acknowledgements

The authors express their gratitude to all of Covalion, especially Dr. Martin Glückler, for their support and contributions throughout this study. Their

sharing of expertise and experience made this study possible and made progress. We are particularly grateful for their help with data collection, technical discussions, and critical feedback that greatly improved the quality of this research. Their commitment to innovation and collaboration was an inspiration, and this work would not have been possible without their support.

References

[1] RK Mobley, An Introduction to Predictive Maintenance, 2nd ed. Burlington: Butterworth-Heinemann, 2002.

[2] T. Zonta, CA Costa, RR Righi, MJ de Lima, ES da Trindade, and GP Li, "Predictive maintenance in the Industry 4.0: A systematic literature review," Computers & Industrial Engineering, volume 150, 106889, 2020.

[3] M. B. Jones et al., "ONNX Runtime: Cross-platform, high performance ML inferencing," Proc. MLSys, 2023.

[4] M. Groshev, C. Guimarães, J. Martín-Pérez, and A. de la Oliva, "Toward intelligent cyber-physical systems: Digital twin meets artificial intelligence," IEEE Communications Magazine, vol. 59, no. 8, p. 14–20, 2021.

[5] F. Tao, B. Xiao, Q. Qi, J. Cheng, and P. Ji, "Digital twin modeling," Journal of Manufacturing Systems, vol. 64, pp. 372–389, 2022.

[6] H. Huang, L. Yang, Y. Wang, X. Xu, and Y. Lu, "Digital Twin-driven online anomaly detection for an automation system based on edge intelligence," Journal of Manufacturing Systems, vol. 59, p. 138–150, 2021.

[7] NY An et al., "Digital Twin-Based Hydrogen Refueling Station (HRS) Safety Model: CNN-Based Decision-Making and 3D Simulation," Sustainability, vol. 16, no. 21, p. 9482, 2024.

[8] X. Yu, X. Yang, Q. Tan, C. Shan and Z. Lv, "An edge computing-based anomaly detection method in IoT industrial sustainability," Applied Soft Computing, vol. 128, 109486, 2022.

[9] AIOTI WG Energy, "Edge driven Digital Twins in distributed energy systems," White Paper, January 2024.

[10] J. Protner et al., "Edge Computing and Digital Twin Based Smart Manufacturing," IFAC-PapersOnLine, vol. 54, no. 1, p. 831–836, 2021.

[11] MA Belay et al., "Unsupervised Anomaly Detection for IoT-Based Systems: A Review," Sensors, 2023.

[12] J. Navajas et al., "A comprehensive analysis of hydrogen refueling station incidents: Unveiling contributing factors," Reliability Engineering & System Safety, 2025.
[13] Y. Suzuki et al., "Machine learning model for detecting hydrogen leakage in high-pressure lines," PHMAP Proceedings, 2023.
[14] B. Chizubem et al., "Real-time monitoring using digital platforms for enhanced operations in hydrogen facilities," International Journal of Hydrogen Energy, 2025.
[15] Google AI Edge, "Model optimization with LiteRT (quantization & pruning)," 2024.
[16] MA Hasanpour et al., "EdgeMark: An automation and benchmarking system for optimized On-Device AI," Journal of Systems Architecture, 2025.
[17] D. Lane et al., "Benchmarking low-power machine learning systems," Proc. MLSys Workshops, 2022.
[18] R. van Dinter et al., "Predictive maintenance using digital twins: A systematic literature review," Information & Management, 2022.
[19] S. Heydari et al., "Tiny Machine Learning and On-Device Inference: A Survey," ACM Computing Surveys, 2025.
[20] M. Hermansa et al., "Sensor-Based Predictive Maintenance with Reduction of False Alarms," Sensors, 2021.
[21] MDR Kabir et al., "Digital Twins for IoT-Driven Energy Systems: A Survey," IEEE Access, 2024.
[22] H. Zhang, S. Gupta, and K. Keutzer, "Accelerating deep learning inference via model quantization,", IEEE Micro, vol. 41, no. 3, pp. 20–28, May–Jun. 2021..
[23] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint* arXiv:1503.02531, 2015.
[24] Z. Tian, H. Lv, W. Zhou, C. Zhang, and P. He, "Review on equipment configuration and operation process optimization of hydrogen refueling stations," *International Journal of Hydrogen Energy*, vol. 47, no. 5, pp. 3033–3053, 2022.
[25] R. van Dinter, B. Tekinerdogan, and C. Catal, "Predictive maintenance using digital twins: A systematic literature review," *Information and Software Technology*, vol. 151, p. 107008, 2022.