Generative Digital Twin Framework for Reliable and Robust AI-Powered Prognostic Systems

Zafer Yiğit
Volvo Construction Equipment - Sweden,
Mälardalen University - Sweden; email: zafer.yigit@mdu.se

Håkan Forsberg
Mälardalen University - Sweden ; email: hakan.forsberg@mdu.se

Masoud Daneshtalab
Tallinn University of Technology - Estonia,
Mälardalen University - Sweden; email: masoud.daneshtalab@mdu.se

Abstract

Reliable operation of heavy-duty construction equipment is essential for maintaining productivity and minimizing maintenance costs. However, developing robust Prognostics and Health Management (PHM) systems is challenging due to limited labeled failure data. This paper investigates an engine-focused digital twin framework that combines machine learning, physics-informed modeling, and synthetic data generation. The approach leverages a digital twin to generate physically consistent datasets using control-driven inputs and physics-based system responses. Previous work demonstrated the feasibility of data-driven PHM on real engine data, but limited data coverage restricts generalization. The proposed framework enables scalable generation of diverse operating conditions. Early results indicate improved robustness and prediction stability compared to purely data-driven models.

Keywords: Prognostics and Health Management, Machine Learning, Physics-Informed, Generative Models, Digital Twin.

1 Machine Learning Approaches for PHM in Engine Systems

Unexpected failures in heavy-duty construction equipment can lead to costly downtime, reduced work efficiency, and increased maintenance expenses. Conventional diagnostic systems mainly rely on fault codes defined in standards such as J1939 [1]. These systems typically detect failures only after they occur and therefore cannot prevent long-term damage. In addition, Diagnostic Trouble Codes (DTCs) provide only a general description of the fault, making root-cause identification and diagnosis difficult and time-consuming.

Prognostics and Health Management (PHM) systems aim to address this limitation by continuously monitoring machine health and predicting potential failures before they occur. In recent years, machine learning methods have shown promising performance for anomaly detection, fault classification, and remaining useful life (RUL) estimation in complex industrial systems.

In our previous work, machine learning and deep learning approaches were applied to real sensor data collected from controlled diesel engine experiments focusing on airpath systems. The study investigated two common failure types: boost air leakage and exhaust gas recirculation (EGR) clogging. Multiple models including Random Forest, XGBoost, convolutional neural networks, and recurrent neural networks were evaluated for anomaly detection and failure classification. The results indicate that ensemble methods such as XGBoost and recurrent models such as LSTM perform effectively in anomaly detection and fault classification tasks [2], demonstrating the feasibility of data-driven PHM approaches in practical industrial settings.

However, these results also highlight a key limitation of purely data-driven approaches. In real-world applications, failure data typically represents only a small fraction of the overall dataset, and reproducing such failures in controlled environments is costly. As a result, models trained only on real data may struggle to generalize to unseen operating conditions.

2 Digital Twin and Physics-Informed Modeling for Synthetic Data Generation

To address this challenge, our research investigates the use of engine digital twin frameworks to generate synthetic data that can support the training of robust PHM models. The architecture of the proposed digital twin framework is illustrated in Figure 1.

A generative scenario model produces diverse engine operating cycles, which are used as inputs for the virtual engine management system (vEMS) and the engine control unit. The vEMS represents the engine control software together with calibration datasets and determines actuator commands and control signals during the simulation. The input signals

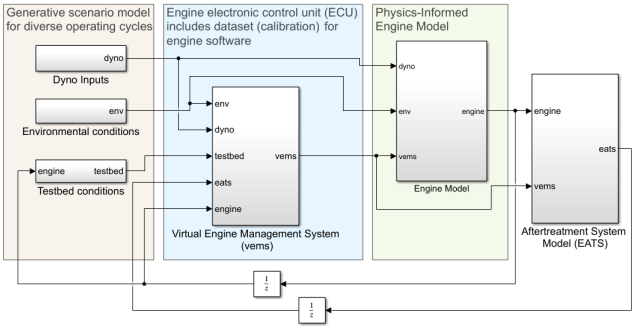

Figure 1: Engine Digital Twin Framework for Synthetic Data Generation in PHM Systems

correspond to dynamometer-controlled (dyno) operating conditions (e.g., engine speed and load), ensuring realistic system excitation under representative operating conditions.

The digital twin is implemented using a simulation-based engine model integrated with the vEMS, enabling a closed-loop interaction between control logic and physical system dynamics. These control signals are then processed by the physics-informed engine model, which combines physical engine dynamics with data-driven components to improve prediction accuracy. The resulting outputs are further processed by the aftertreatment system model (EATS), which is based on established simulation models used in engine system analysis and provides a consistent representation of emission-related processes.

This integrated framework enables the generation of physically consistent synthetic datasets that reflect realistic operational scenarios by combining control-driven inputs with physics-based system responses. Such datasets can augment real-world data and improve the robustness of anomaly detection, fault classification, and RUL prediction models. In particular, the framework supports scalable data generation by producing diverse operating conditions beyond those observed in available real-world datasets.

Ensuring that the generated synthetic data remains representative of real-world behavior is a key challenge and is considered part of the ongoing validation process. Initial validation studies indicate consistent trends between the simulated and measured system responses, supporting the realism and practical applicability of the generated synthetic data.

Physics-informed machine learning has recently emerged as an important approach for improving the robustness and interpretability of data-driven models in monitoring and anomaly detection applications [3].

As part of ongoing research, physics-informed recurrent architectures are being developed to further strengthen the predictive capability of the digital twin. In this approach, thermodynamic energy balance equations are embedded directly into the state evolution of recurrent neural networks. Similar approaches have been explored for thermal system dynamics [4], while our recent work further investigates embedding thermodynamic dynamics into recurrent architectures to improve robustness [5].

Future work will explore advanced generative modeling approaches to further increase dataset diversity. In particular, transformer-based diffusion models will be investigated to generate diverse operating scenarios, which can be used as inputs to the digital twin to produce physically consistent sensor data for PHM model training.

The long-term objective is to develop a robust digital-twin-supported PHM framework that enables reliable anomaly detection and prognostic modeling for heavy-duty construction machinery by combining machine learning, physics-informed modeling, and generative scenario modeling.

References

[1] M. R. Stepper, S. R. Butler, and G. G. Zhu, “On-board diagnostics, a heavy duty perspective,” *SAE Technical Paper Series*, vol. 95147, 1995.

[2] Z. Yigit, H. Forsberg, and M. Daneshtalab, “Machine learning-based prognostic approaches for construction equipment powertrain systems,” *IEEE Intelligent Vehicles Symposium (IV)*, 2025.

[3] Y. Wu, B. Sicard, and S. A. Gadsden, “Physics-informed machine learning: a comprehensive review on applications in anomaly detection and condition monitoring,” *Expert Systems With Applications*, vol. 255 (2024) 124678, 2024.

[4] X. Tang, Q. Hong, Y. Liu, B. Wang, Z. Cui, and W. Shao, “Physics-informed LSTM based dynamic model of heat exchanger and application on the thermal management system considering time delay,” *International Journal of Heat and Mass Transfer*, vol. 250 (2025) 127322, 2025.

[5] Z. Yigit, M. Daneshtalab, and H. Forsberg, “Physics-informed recurrent architecture with embedded thermodynamic dynamics for robust sequence modeling,” *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Proc. ESANN 2026 (accepted).