

Machinery health monitoring system

Lakshmi suvarna

School of Computing and
Information Technology, REVA
University, Bangalore, India
suvarnakalluru@gmail.com

Y.Venkateswarlu

School of Computing and
Information Technology, REVA
University, Bangalore, India
yeruvavenkatesh15@gmail.com

Abhiram Sai

School of Computing and
Information Technology, REVA
University, Bangalore, India
kemparjabhiram@gmail.com

Danish Mohan

School of Computing and
Information Technology, REVA
University, Bangalore, India
danishnaik8243@gmail.com

Dr. Ravi Shankar

School of Computing and
Information Technology, REVA
University, Bangalore, India

Abstract— The present paper introduces the concept of Machinery health monitoring system AI: predictive maintenance software-defined digital twin hardware-independent framework on industrial machinery. Specifically, the machine model consists of a discrete-time Markov Chain with three machine operating states: healthy, degrading, failure that allows realistic generation of synthetic data without the use of physical sensors. Multivariate time-based series data, including temperature, vibration and load, is also enhanced by stochastic perturbation in order to simulate the real-life operating environment. A sequence-to-sequence LSTM autoencoder using a temporal attention mechanism on which an unsupervised anomaly detecting model is trained is used to learn normal operational patterns. The model calculates reconstruction error as score of an anomaly which can be used to detect deviation to healthy behavior reliably. An automatic thresholding method allows the machine states to be classified in real-time. This system can run at a 1 Hz streaming rate and sub-milliseconds inference latency on a GPU system. The architecture proposed will consist of FastAPI-based back-end as a server, WebSocket layer as a means of communication, and a React-based front-end as a visualization and monitoring tool. This study shows that it can be scaled and repeated to predictive maintenance with synthetic data, and this provides the basis to make the transition to actual industrial applications.

Keywords— Predictive Maintenance, Digital Twin, LSTM Autoencoder, Temporal Attention, Anomaly Detection, Markov Chain, Time-Series Analysis, Synthetic Data, Real-Time Systems, Industrial AI.

I. INTRODUCTION

The growing complexity of the modern industrial systems has increased the necessity to have smart maintenance tactics that guarantee reliability, safety, and operational efficiency. Conventional maintenance methods such as reactive and preventive maintenance are usually ineffective in responding to the unforeseen failures and maximize the use of resources. Preventive maintenance can result in unnecessary servicing on equipment that is in good operating condition whereas reactive maintenance will cause unplanned down time and a high cost of repair. Thus, predictive maintenance has become a data-oriented

solution that allows identifying faults in time and making decisions.

Traditional predictive maintenance systems are very dependent on physical sensors, embedded devices, and Internet of Things (IoT) infrastructure to monitor real-time information of the machinery. Nonetheless, these solutions present big costs of deployment, dependencies on hardware and restrictions to scale especially in applications where instrumentation is limited or infeasible. The current techniques are thus reliant on hardware based data collection and are therefore not flexible or easy to access. To fill this void, this paper introduces a predictive maintenance system named Machinery health monitoring system AI which is entirely a software-implemented framework that removes the need to incorporate physical sensing into the framework by a digital twin paradigm.

The suggested system models the behaviour of the machine by means of a probabilistic discrete-time Markov Chain, creating realistic multivariate time-series data as one of the possible operational states. They use an unsupervised anomaly detection model that is made out of a sequence-to-sequence LSTM autoencoder with temporal attention that is trained to learn normal operation patterns and detect deviations when it uses reconstruction error (mean squared error) as an anomaly score. The system uses 1 Hz real-time simulation rate with low-latency inference with the assistance of the use of GPUs. WebSocket-based communication and the integration of a real-time processing pipeline allow maintaining the health of the machines and visualizing it in real time.

The article presents an independent hardware-based predictive maintenance system which uses probabilistic simulation and deep learning to detect anomalies. The offered solution offers a flexible and scalable framework, which can be implemented to various industrial sources which prove that it can be regarded as an alternative to monitor systems that are sensor-dependent.

II. LITERATURE SURVEY

A. Digital Twin-Driven Predictive Maintenance

Digital twin is the new paradigm of predictive maintenance because it provides the possibility of virtual representation of physical systems and their monitoring and analysis. Ismail et al. [1] provide an in-depth survey of the digital twin-based predictive maintenance with a focus on its ability to model the real-time machine behavior and provide the data-driven decision-making. In the same manner, Chen et al. [2] comment on the use of machine learning in the context of digital twins and show that predictive models increase the quality of fault detection and efficiency.

Abd Wahab et al. [3] also indicate the issues of scalability and data dependency as the main problems of current digital twin-based maintenance systems, especially in an industrial setting where the data originates are heterogeneous. Van Dinter [4] puts emphasis on the architectural design issues, and the need to consider modular and decoupled systems structure. Although these studies prove the efficiency of digital twins in predictive maintenance, one of the main limitations that can be seen in all these works is that they largely rely on sensor data in the real world in order to train the model and ensure that it is valid. This dependency poses limitations to cost of deployment, accessibility and scalability.

Moreover, the vast majority of the currently used frameworks have hybrid designs that combine physical sensing with simulation, and this constrains their usage in the purely software-defined setting. Conversely, the proposed work seals this divide with the introduction of a fully software-based digital twin, which can produce realistic synthetic data streams, which are not dependent on hardware but yet the fidelity of systems.

B. LSTM Autoencoder-Based Anomaly Detection

The use of deep learning models, time-series data in anomaly detection the deep learning model, in particular, LSTM autoencoders (which capture the temporal associations), have been popular in time-series data because they can capture the temporal association. Bamboula et al. [5] show that the LSTM autoencoders can be effective in terms of modeling normal operating conditions and identifying anomalies with reconstruction error without use of labeled failure data. Likewise, frameworks of real-time anomaly detection based on LSTM autoencoders are studied in [6] where a stream of data is handled in real-time to spot deviations occurring in dynamical circumstances. Sharma and Gupta [7] also support the relevance of the unsupervised learning methods in industrial settings where limited data on faults is available.

Nevertheless, one of the main weaknesses of traditional LSTM autoencoder models is the fact that they cannot localize anomalies in a time sequence with high accuracy because they use the compact latent representations that can hide the important variations at the timestep level. Also, the current literature lacks adequate coverage of interpretability that is

critical in making industrial decisions. The standard LSTM autoencoders have powerful temporal modelling compared to the others including variational autoencoders or transformer-based modelling, however, they do not have special mechanisms to identify the most significant parts of the input sequence.

To overcome these drawbacks, the proposed work will combine a temporal attention mechanism, which allows the model to weigh the importance to each timestep, and better detect the non-sudden patterns of degradation.

C. Integration of Digital Twin and Deep Learning

Digital twin implementation in conjunction with deep learning has received considerable interest as a way to resolve the issue of data shortage and enhance the prediction performance. Karakia et al. [8] suggest simulated data-driven predictive analytics are proposed through a machine learning-based digital twin framework, which shows how simulation and learning could be combined. On the same note, Ma et al. [9] present an extensive overview of the AI-controlled digital twin frameworks, where they note that they can be used to undertake intelligent monitoring and adaptive decision-making.

Hybrid systems, which integrate a simulation environment and machine learning models, are also further discussed in [10] where predictive abilities are promoted by involving virtual and real-world information. The unsupervised methods of learning combined with the digital twins are studied in [11], where it is revealed that the work of anomaly detection in the absence of labeled datasets can be conducted.

Nevertheless, the omnipresence of these advancements is that, in all these studies, one of the limitations is the use of partially real data or hybrid data pipelines, which remain limiting to scalability and flexibility of deployment. Also most of the existing systems lack support for real time streaming demands; hence their usage in continuous monitoring is restricted. The proposed system will be unique, as it creates a fully synthetic, probabilistically driven data pipeline with an added anomaly detector that is based on deep learning, producing a complete end-to-end software solution with real-time performance.

D. Advanced Predictive Maintenance Structures and Research Views.

Recent developments on predictive maintenance have been on the enhancement of model robustness, scaling and computational performance. Hu et al. [12] propose a parallel LSTM autoencoder based failure prediction model that showed better results because of the optimization of the architecture. Chen et al. [13] expand on the wider role of digital twins in predictive maintenance with a particular focus on the fact that they make it possible to monitor the system continuously and have adaptive behavior. The principal idea of digital twins, which is explained in [14], only supports their significance in contemporary industrial ecosystems.

Nevertheless, one weakness that is realized critically in these methods is that they are costly to deploy in terms of hardware infrastructure and integration overheads. Numerous solutions that exist are also not streamlined with real time communication mechanisms that are needed to provide responsive monitoring solutions.

The suggested work will solve these issues with the help of a lightweight, hardware-independent framework that can be used with a real-time streaming architecture. The system allows low-latency processing and real-time monitoring without the physical sensing infrastructure by using the system based on WebSocket-based communication and the use of GPUs to perform the inference. This makes the suggested method a viable and available alternative to the traditional predictive maintenance systems.

Altogether, the existing literature points to the drastic shift towards the combination of simulation, deep learning, and real-time systems. The given system develops these directions to come up with a unified framework integrating probabilistic modeling, synthetic data generation, and attention-based anomaly detection.

III. METHODOLOGY

A. Probabilistic Data Generation of Digital Twin.

The offered system is grounded on a full software-defined digital twin framework, which emulates the working behavior of an industrial machine without the employment of physical sensors or IoT devices. A discrete-time Markov Chain (DTMC) promotes the data generation process and characterizes the machine through three operating states, including HEALTHY, DEGRADING, and FAILURE.

In Markov engine, the Markov engine calculates the state of the system at every timestep given a predetermined transitions probability structure. The model is based on the memoryless property, which means that the following state is determined by the current state. The normal state is called HEALTHY, progressive wear is called DEGRADING and the non-recoverable condition is called FAILURE. The failure state is considered an absorbing state, so once the system gets on this state, it would be in that state until a manual reset. After the state determination, the Sensor Simulator produces the values as baseline values of the temperature, vibration and load according to the state in hand. The values provided as the baseline are meant to be realistic, i.e. the machine behavior of degradation resulting in more temperature, vibration and load.

In order to be more realistic, a Noise Injector module is added which adds stochastic variations to the signals. This is modeled with Gaussian noise to model measurement error, random walk drift to model long-run changes, and injection of spikes in case of failure to model unforeseen abnormal phenomena. The resulting output is an on-going realistic time-series stream of data at a constant rate of 1 Hz.

B. Feature Engineering and Data Preprocessing

The sensor data generated undergoes a feature engineering pipeline to generate a structured representation to be used as an input to model. The workings of every timestep are modeled as a nine-dimensional feature vector based on three major signals, namely temperature, vibration, and load.

The feature set is made up of three classes namely raw sensor values, first-order difference (delta) features, and rolling standard deviation features. The uncooked values are the immediate sensor values. The delta features capture the information between the successive timesteps such that the model is able to capture differences and trends in the data. The rolling standard deviation is calculated on a small window in order to represent a local variability and signal changes. Every feature is normalized with MinMaxScaler which is fitted only on the data on the healthy operations. This makes sure that the distribution of normal behavior is learnt by the model and that when inference is being performed, it is possible to find deviation of the normal behavior. The data after the processing is arranged in 50 timestamp long sequences through a sliding window method. Also, an 80 timestep rolling buffer is stored to allow real-time processing.

C. LSTM Autoencoder with Temporal Attention for Anomaly Detection

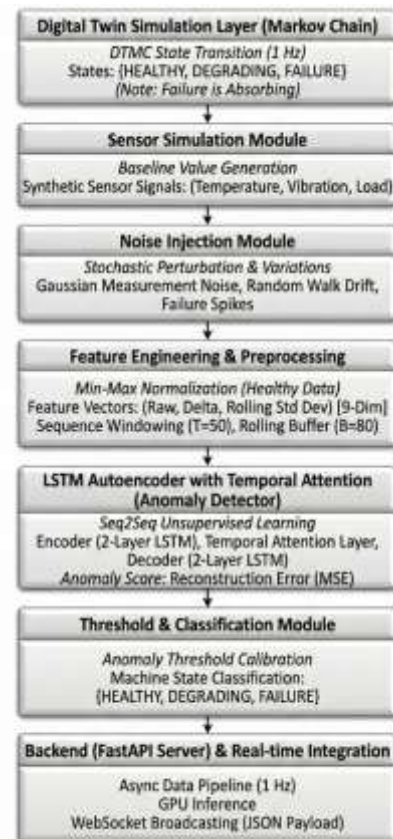


Figure 4.1: System Architecture / Pipeline Diagram

Anomaly detection part is fulfilled with the help of sequence-to-sequence LSTM autoencoder with the addition of temporal attention mechanism. This model is meant to acquire the patterns of normal machine running and detect anomalies on the basis of reconstruction error. The encoder is an LSTM network with two layers which is fed on input sequences of length 50 and with nine features/timestep. The model first derives a time representation of the input data and then processes the time representation with an attention mechanism which attributes significance to various timesteps in the sequence. This gives the model the ability to concentrate on the most relevant parts of the input in coming up with the reconstructed output.

The coded representation is condensed into a latent space and the coded message is then fed to the decoder which reinvents the original input sequence through the two-layered LSTM network with a linear output layer. The model is trained with the help of data that is related to the normal functioning of machines. In the process of inference an anomaly score is calculated as the difference between the input sequence and the reconstructed sequence. Greater error in reconstruction means an abnormality of behavior, associated with deteriorating or broken states. The model is enhanced by the use of temporal attention to detect the instances where an anomaly takes place in the sequence.

D. Real-Time Inference Pipeline and System Integration

The system works as a real time pipeline and combines data generation, preprocessing, model inference and visualization. This is carried out at a fixed frequency of 1 Hz and with the help of an asynchronous backend. The system will at every timestep perform State transition, sensor simulation, noise injection, feature extraction and model inference respectively.

The trained LSTM autoencoder processes the processed data and the resulting reconstruction error is taken as the anomaly score. This score is classified through a threshold-based classification mechanism that assigns this score to discrete health states; HEALTHY, DEGRADING, or FAILURE. FastAPI is used to implement the backend, which uses the WebSocket interface to communicate with the front end. Every timestep, a structured JSON message with sensor measurements, anomaly index, and anticipated health condition is sent to attached clients.

Front end constitutes React and TypeScript in which real time visualization of incoming data is in the form of charts, gauges and neural network representation. By using high-performance computers and the acceleration of its system on the basis of a graphical processor, the system displays the health of the machine in real-time and allows customers to view the results immediately. This end-to-end predictive maintenance system is fully integrated in software utilizing this integrated pipeline.

IV. RESULTS AND ANALYSIS

A. Sensor Data Simulation and Behavioural Analysis

The assessment of the suggested system will be initiated by the analysis of the synthetic data produced by the digital twin framework. The ability of the system to recreate realistic machine behavior is shown by the real-time visualization of the

temperature, vibration, and load parameters. In the normal working conditions, all three signals have continuous and periodical variations having slight stochastic fluctuations. The variations are brought about by the components of Gaussian noise as well as drift that are introduced in the simulation process which ensures that the produced data is highly similar to the real-life sensor data.

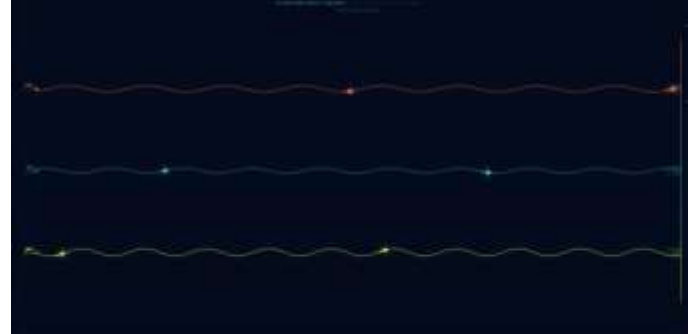


Fig. 5.1: Real-time sensor signal visualization (temperature, vibration, and load)

As can be seen in Fig. 5.1, the temperature signal has a steady oscillatory profile, but both vibration and load signals are within reasonable operation limits. Lack of sudden spikes or defective patterns gives certainty that the system is indicative of healthy conditions of machines. This consistency is important as it would be used to train the anomaly detection model, and it would provide a reliable baseline of normal behaviour.

As the system passes through the degrading state, the signal characteristics will be deviated gradually. Temperature and vibration signals have a small upward trend meaning that there is an upward trend of mechanical stress, and load variations are related to decreased efficiency of the system. These modifications are mild and spread out throughout time and thus they are hard to detect through the traditional ways of threshold. This gradual degradation patterns can be captured by the simulation, and this demonstrates the strength of the probabilistic modelling and noise injection processes.

B. Anomaly Detection Performance and Model Interpretation

The LSTM autoencoder with the temporal attention is used to test the system in terms of the anomaly detection ability. The model is only trained using healthy data, which allows the model to learn normal temporal dependencies between the sensor signals. In the inference phase, the model reinvents the input sequences and calculates reconstruction error which is used as the score of anomalies. The error in reconstruction is also consistently low in the healthy state, which means that input sequences are correctly reconstructed. The system progresses into the degrading state where there is gradual rise in the anomaly score. It is a sensitive behaviour that shows how sensitive the model is to the slight variations and how it can be used to identify the anomaly at an early stage before total failure takes place.

Fig. 5.2 presents the visualization of the neural network structure and neural network activations. The temporal attention mechanism identifies timesteps that have the biggest contribution towards the anomaly score. This enhances the interpretability as the system is able to determine points of deviation within the sequence. The model is not entirely dependent on the worldwide chain of sequences but rather it concentrates on key time zones which makes it more effective in the detection.

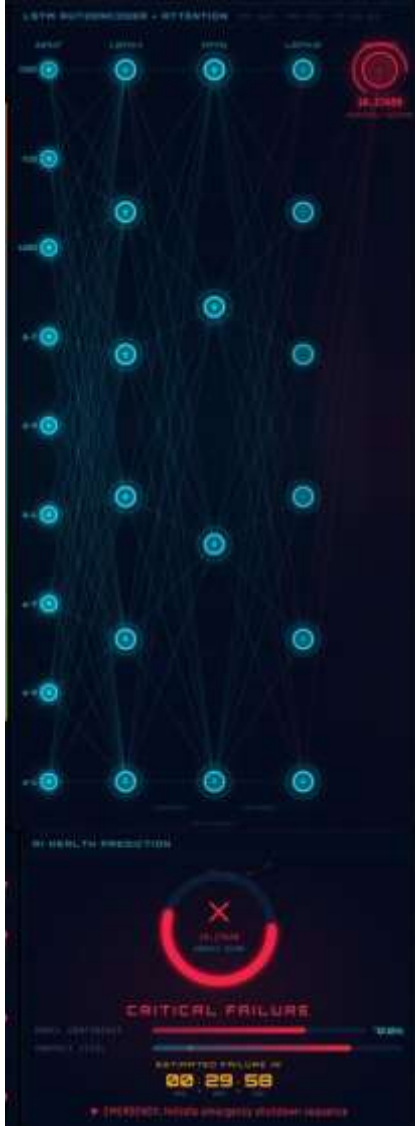


Fig. 5.2: LSTM autoencoder with temporal attention visualization

In the state of failure, the score of anomaly rises significantly because it contains spike injections and unusual signal patterns. This distinction between normal and faulty states is an indication of a well-developed detection performance. This model exhibits a stable behaviour without erratic changes which makes detection of anomalies in a steady and reliable manner.

C. Real-Time System Performance and Visualization

The suggested system is used in a real time system and the fixed update rate is 1 Hz. The approach of asynchronous pipeline implementing data generation, preprocessing and inference guarantees the continuous and efficient functioning. The way it has used the GPU acceleration is that it can inference models very fast and therefore the system can process incoming data without any latency problem.

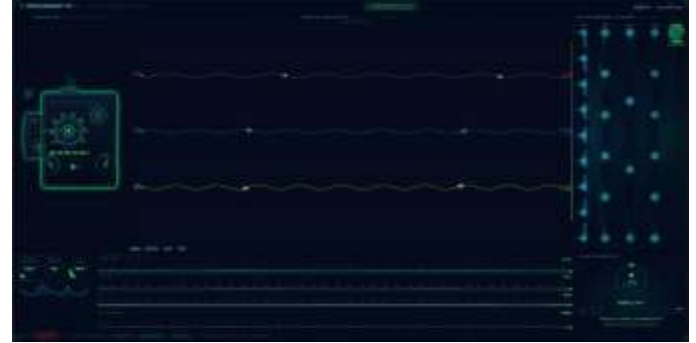


Fig. 5.3: Real-time monitoring dashboard interface

Fig. 5.3 shows the entire monitoring dashboard that is a composite of various visualization elements. The system offers time series graphs on sensor signal, anomaly scores indicators, and machine health. In the case of real-time monitoring, monitoring delays are removed because the WebSocket-based communication will send the updates immediately between the backend and the frontend. The dashboard design makes it easier to use since it displays the information at varying levels. The high-level indicators allow seeing health conditions of the machine instantly, whereas detailed charts can help to analyse the behaviour of the signals in more detail. Transparency can also be enhanced by the fact that, in addition to other items, the visualization of neural networks can be used to demonstrate the way the model processes input data. The system has stable performance of all the components and no delays and disruption witnessed during the continuous use. It shows that the system architecture is effective in the application of real-time predictive maintenance

D. System Effectiveness, Failure Analysis, and Limitations

The functionality of the proposed system of AI in general and the machinery health monitoring system can help to determine its effectiveness as it can easily detect and identify machine health. The system is able to determine the transitions between the condition of health to the condition of degradation and ultimately failure. The anomaly detection module generates systematic and understandable results that allow critical states to be detected on time. As illustrated in Fig. 5.4, the system identifies the failure conditions that significantly score high in terms of anomaly, which causes a condition of Critical Failure classification. Other indicators that can be used to support predictive maintenance decisions have also been included in the visualization like the level of anomaly and estimated time-to-failure. It shows that the system is not just able to identify the faults, but also offer actionable information.

One of the strong points about the system is the design of the system which is independent of hardware. Using the purely simulated data, the system will not have to use any physical sensors and makes its deployment very complex. This renders it extremely scaled and research and prototyping ready. Nonetheless, some shortcomings should be taken into account. The system relies on artificial data, and therefore, its functionality with respect to real world performance is determined by how well the simulation model is developed. Also, the anomaly detection model is reliable, and additional testing on real-life industrial data is required to make it robust and generalized.

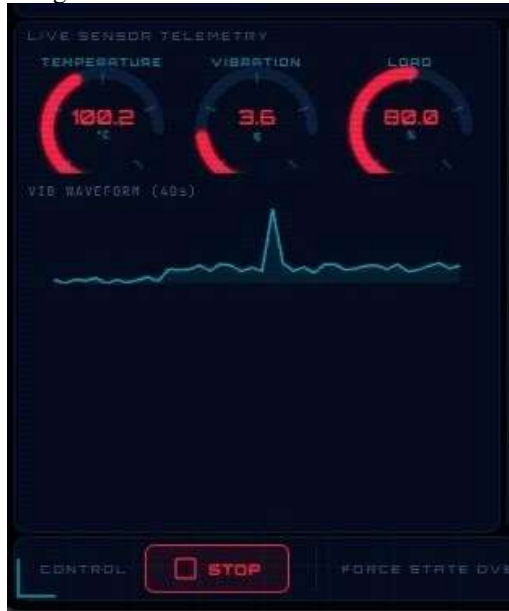


Fig. 5.4: Anomaly detection output showing critical failure state

Irrespective of these constraints, the proposed system shows high and viable nature of predictive maintenance. The combination of the digital twin simulation and deep learning-based anomaly detection with the real-time visualization will offer an effective solution to the issue of machine health monitoring.

V. CONCLUSION

This publication introduced the machinery health monitoring system AI, which is a complete software-based predictive maintenance system that does not rely on physical sensors with the use of a digital twin. The system combines probabilistic data generation with discrete time Markov Chain with an LSTM auto-encoder based anomaly detection model augmented with temporal attention. The system is capable of detecting deviations of the normal machine behavior by utilizing reconstruction error using the multivariate time-series data. The entire pipeline that also includes data simulation, feature engineering, real-time inference, and visualization is an illustration of a scalable and coherent architecture that can be continuously monitored. The deployment demonstrates the practicability of the implementation of predictive maintenance capability fully within software with the non-dependence on external hardware infrastructures. The next step in work will be

to expand the system to real-life application, including live sensor data, to the already existing digital twin system. The further enhancements comprise the model modification to suit different types of machines, the increase in its resistance to different operating conditions, and the improvement of its ability to detect anomalies. More detailed investigation of high-level deep learning models and the optimization of the real-time stream processing will also be taken into account. The developments are geared toward increasing the scalability aspect and multi-machine monitoring to allow wider application to the industrial setting. The next step in work will be to expand the system to real-life application, including live sensor data, to the already existing digital twin system. The further enhancements comprise the model modification to suit different types of machines, the increase in its resistance to different operating conditions, and the improvement of its ability to detect anomalies. More detailed investigation of high-level deep learning models and the optimization of the real-time stream processing will also be taken into account. The developments are geared toward increasing the scalability aspect and multi-machine monitoring to allow wider application to the industrial setting.

REFERENCES

- [1] R. van Dinter, B. Tekinerdogan, and C. Catal, "Predictive maintenance using digital twins: A systematic literature review," *Information and Software Technology*, vol. 150, 2022.
<https://www.sciencedirect.com/science/article/pii/S0950584922001331>
- [2] I. Mavrelis, A. Chatzopoulos, and P. Zacharia, "Hybrid deep learning for predictive maintenance in industrial machinery using LSTM and MLP models," *Machines*, vol. 14, no. 2, 2026.
<https://www.mdpi.com/2075-1702/14/2/191>
- [3] H. Wei et al., "Integrating digital twin technologies for maintenance 4.0: A review," *Energy and Buildings*, 2025.
<https://www.sciencedirect.com/science/article/pii/S0360132325014672>
- [4] A. Schwarz et al., "Data augmentation in predictive maintenance applicable to industrial systems," *Artificial Intelligence Review*, 2024.
<https://link.springer.com/article/10.1007/s10462-024-11021-9>
- [5] "LSTM-based autoencoder models for real-time anomaly detection," *Journal of Hydroinformatics*, 2024.
<https://iwaponline.com/jh/article/26/2/441/99926>
- [6] "Digital twin applied to predictive maintenance for Industry 4.0," 2024.
https://www.researchgate.net/publication/381933841_Digital_Twin_applied_to_Predictive_Maintenance_for_Industry_40
- [7] U. Mutlu et al., "A systematic review of AI, BIM, and digital twins for predictive maintenance," *Buildings*, 2026.
<https://www.mdpi.com/2412-3811/11/3/87>
- [8] K. A. Kushal and F. Gueniat, "AI-enhanced IoT systems for predictive maintenance using digital twin," *arXiv preprint*, 2025.
<https://arxiv.org/abs/2511.12175>
- [9] L. Ismail et al., "A systematic review of digital twin-driven predictive maintenance in industrial engineering," *arXiv preprint*, 2025.
<https://arxiv.org/abs/2509.24443>
- [10] V. Karkaria et al., "A digital twin framework utilizing machine learning for predictive maintenance," *arXiv preprint*, 2024.
<https://arxiv.org/abs/2408.06220>
- [11] E. G. Kaigom, "Deep generative digital twin with variational autoencoder for predictive monitoring," *arXiv preprint*, 2025.
<https://arxiv.org/abs/2509.12740>
- [12] "Artificial intelligence-driven fault detection and predictive maintenance systems," *Engineered Science*, 2023.



<https://www.espublisher.com/journals/articlehtml/engineered-science/10.30919-es2061>

- [13] "Prospects for predictive maintenance and intelligent industrial systems," ACM Digital Library, 2026.

<https://dl.acm.org/doi/10.1145/3779475.3779525>

- [14] "Lifelong updating of digital twin models using autoencoder and LSTM," 2026.[link](#)