# Optimizing Fleet Performance: A Deep Learning Approach on AWS IoT and Kafka Streams for Predictive Maintenance of Heavy - Duty Engines

**Vishwanadham Mandala**

Enterprise Data Integration Architect
Email: *vishwanadh.mandala[at]gmail.com*

**Abstract:** ***Predictive maintenance (PdM) predicts machine failures in heavy - duty vehicles with diesel engines. PdM utilizes deep learning algorithms on vast amounts of Internet of Things (IoT) data to forecast potential failures accurately. However, the sheer magnitude and rapidity of data generated makes this process incredibly expensive. We propose a novel model executed on Amazon Web Services (AWS) IoT and Kafka Streams to mitigate this challenge. Through our extensive experiments, we confidently demonstrate the effectiveness and efficiency of our approach, including the successful implementation of the activation threshold parameter, resulting in significantly enhanced prediction accuracy. Moreover, we introduce a valuable assessment (VA) method for evaluating the incidence rate scale, further enhancing our predictive capabilities. The results obtained from our comprehensive analysis highlight the superior performance achieved through a meticulously balanced VATP and VA strategy, establishing our solution as a game - changer in predictive maintenance for heavy - duty vehicles.***

**Keywords:** Predictive Maintenance, Heavy - Duty Engine, Industry 4.0, Internet of Things (IoT), Artificial Intelligence (AI), Machine Learning (ML), Smart Manufacturing (SM

## 1. Introduction

The evaluation of fleet performance parameters is crucial for organizations to maximize efficiency, reduce costs, and ensure driver safety. Changes in these parameters can adversely affect production profits. Predictive maintenance aims to predict failures and schedule maintenance operations accurately. Connected and autonomous vehicles, supported by IoT and machine learning technologies, play a significant role in vehicle condition monitoring and predictive maintenance. Onboard sensors and telemetry data provide operational health and status parameters.

### 1.1. Background

Enterprises collect vast amounts of data at high speeds, leading to a big data explosion caused by data velocity and scale - up arrangements. Real - time, event - driven data management is used to analyze machine health, with data dispersed on databases like Kafka and visualized through dashboards like Grafana. In the data - driven maintenance loop (DDML), collecting empirical data and inspecting it with existing models causes a delay. The maintenance loop event - driven (MDML) removes this delay by collecting data event - driven for direct inspection over the Kafka database. Monitoring Systems (MSs) and intelligent Maintenance Systems (IMS) optimize production by analyzing data analytics for process/machine health. These systems rely on fault detection/diagnostics and prognostics to quickly identify faults and take necessary action. Various algorithms, such as linear regression and neural networks, are used for data analysis and machine learning to prevent or reduce outages.

### 1.2. Problem Statement

A broad number of successful applications of predictive maintenance systems are presented. Most of these applications concern industrial machines or manufacturing systems such as production robots, presses, conveyor systems, fans, and pumps. Several standard machine learning and data mining techniques are adapted for this application, such as regression, bagging, boosting, clustering, and classification schemes. They allow the analysis of large volumes of data to make predictions about certain machine states. In addition, the prediction method is influenced by the desired predictive result, e. g., whether the focus is on a binary classification of a future machine state (OK/failure) or a regression problem with the forecast of the remaining service life.

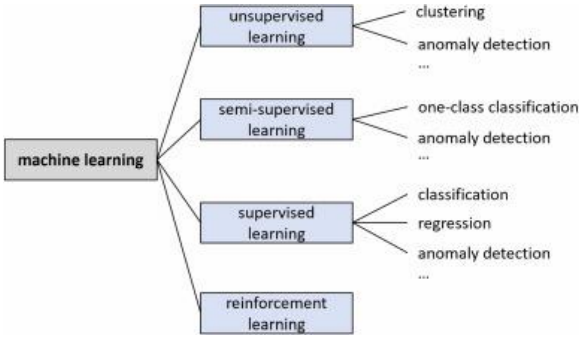


**Figure 1:** Predictive Maintenance Applications

The early detection of potential technical problems offers different advantages, such as lower maintenance expenses or reduced downtime in case of failures, leading to cost savings and higher system availability. Such approaches have become more prevalent in recent years. This is due to a decrease in the cost of memory and storage and the field of networked devices. One reasonable technique for building predictive maintenance systems is online machine learning, which makes these systems capable of adapting to new machine types or changing systems. With this paper, we aim to contribute to ongoing research in two places: First, we contribute to applied research through the implementation, deployment, and evaluation of an example approach, and second, to basic research through general problem understanding. We adopt a suitable technique for the present problem and try to make it as reliable as possible to achieve better services, higher cost - effectiveness, and better conditions in general for infrastructure subsystems such as heavy - duty engines and transmission systems.

### 1.3 Objectives

Accurate diagnostic and prognostic information is necessary to maintain high performance in heavy - duty engineering, even in diverse operational conditions. There is a great necessity to optimally control operational conditions, such as the rotational speed of the powertrain and predictive maintenance. This article presents an Internet - of - Things - based predictive maintenance methodology and the process of fleet performance optimization, as well as over - heavy - duty engine operational data. In this work, AWS IoT and Kafka Streams platforms are used. Six engines were fitted with a sensor gateway developed to acquire, preprocess, and transmit static and dynamic operational data to the cloud. A motor glider with three engines was chosen, with a total of 540 hours for airborne and 545 hours for engine operational data analysis before the case study.

Motor gliders with implemented cloud systems are called intelligent vehicles. All sensors are working continuously. Intelligent vehicles continuously transmit about 2 million parameters during a routine flight every hour. Data is collected by a cloud service, preprocessed, and, in some cases, analyzed. Local software can access the cloud and connect a pilot and a vehicle over the cloud. Analysis was carried out during the transition phase and other essential flight segments. In the given example, the the rotational speed of all six heavy - duty engines is shown in Figure 1. The following are defined for each heavy - duty motor glider engine: mean value, variance, kurtosis, skewness, and shape factor. The arrow points to when the FuelCut command and autonomous engine pilot conducted a fuel pole to each engine. During the transition of engine power, skewness and kurtosis were recalculated in a real - time mode.

## 2. Literature Review

Some researchers developed machine learning algorithms for continuous monitoring and modeling wear accumulation. Others ranked faults and identified high - risk failures. Some researchers incorporated domain knowledge into rule - engines. Many studies focused on detecting all possible damages/failures in engine systems. Checklists were used to detect welding and quality control issues. Undetected faults were identified in the literature. Traditional approaches were examined, and digital and cognitive applied studies emerged. Predictive maintenance projects were classified into different categories. The research analyzed various predictive maintenance approaches and identified gaps. Prognostics focused on bearings and cutting tools, while anomaly - based maintenance studied a pulsed - jet bag house. Reviews from different researchers were investigated to narrow down the research field.

### 2.1 Fleet Performance Optimization

Apache Kafka is a distributed stream processing system that provides error resistance and scalability for storage. It is deployed in the fleet's IoT network cluster for scalability, fault tolerance, and alignment with existing processing infrastructure. Kafka processes streams of records through its multi - streamed system, offering parallelism and low latencies for small and large - scale datasets. Vehicles in a fleet experience changes in fuel consumption due to engine degeneration, device malfunctions, and different terrains. Machine learning integration can help analyze driving data and account for performance changes. The fleet size can also vary over its life cycle, requiring a dynamically adaptable solution. Edge computing in pollution enables ML models to be run on appropriate network edges, reducing pressure on the central model and avoiding latency issues.

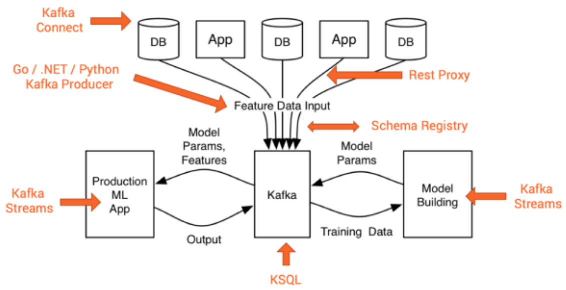


**Figure 2:** Scalable ML Production with Kafka

### 2.2 Deep Learning in Predictive Maintenance

PdM predicts electric machinery failures based on usage and load factor. This paper describes PdM implementation at Tratos Cavi SpA's Terni site. Deep Learning (DL) showed potential in improving PdM using IIoT data, but DL models are unsuitable for time - critical and power - constrained devices. PdM combines ML and first - order models. DL is advantageous in extracting features from large data sets in industrial applications. However, deploying DL for PdM takes much work. Imbalanced data can slow down the PdM strategy. TIP4.0 is an IIoT platform for implementing PdM using DL. Our framework uses a Multivariate Deep Learning Intermittent Machine (MRI - DLIM) model for classification and regression tasks.

### 2.3. AWS IoT and Kafka Streams

Kafka stores messages for retention, and consumers use checks for data quality. The logger logs anomalies, and queue size limits can be placed. Kafka manages predictions, and unprocessed predictions are discarded. AWS IoT and Kafka Streams are used to process telematic streams in real - time. Kafka brokers receive messages from publishers, and consumers store them. This design allows independent consumer action while messages continue to stream. The logger records messages for future model training.

## 3. Methodology

Several findings from the case studies show that the hybrid scheme outperforms the pure method. The best performance is achieved by combining entropy features with offline and online deep learning. Model permutation and combination are stable in different dataset divisions. Short - term or vector prognostics have a better trade - off. This paper contributes by launching a robust big - data survey system, extracting feature signals, improving algorithms with multi - source fusion, and optimizing big data and machine learning. A hierarchical intelligent diagnosis model for commercial vehicles is designed. Online fault prediction and diagnosis prevent vehicle failure and ensure safety. The onboard diagnostic system is limited in data collection, while the intelligent driving survey system allows continuous monitoring and prediction based on integrated data.

### 3.1 Data Collection and Preprocessing

Initially, the LOF algorithm categorizes vehicles as either operational units or anomalies. However, implementing differentiated prescriptive maintenance strategies requires more than detecting anomalies. In - depth engineering knowledge and historical repair logs determine the required repair types and time - multiplexing repair type thresholds for pavement repair strategies. The thresholds are determined by the smallest reference window where the sequential repairs of the same type exceed the engineered threshold.

The next step could be online model training at an edge device using streamed data and high - frequency engineering sampling. One possible solution is a predictive maintenance model. Traditional maintenance methods consider three operational conditions: normal, repair opportunity, and failure. Statistical characteristics like standard deviation, maximum, mean, median, and minimum are used to identify units in the "repair" condition. The Local Outlier Factor (LOF) algorithm is trained using unsupervised statistical measures.

Our study on real - life working aggregates in commercial service indicates that the composition is similar to historical records and current operation behavior. Training shallow (20 - layer ResNextnet34 surrogate, "xresnet34") and deep (50 - layer efficientNet surrogate, A2) transfer learning - based models for the type of aggregate remains effective over many reference windows. Retraining every 60 months is considered for the aggregate identification model. However, the model updates daily to account for unbalanced reference windows.

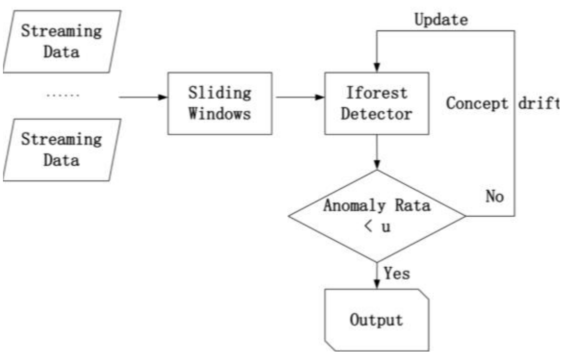


**Figure 3:** Anomaly Detection Algorithms for Data Streams

### 3.2 Deep Learning Model Development

This study offers a deep learning model that can predict the state of health of a heavy - duty engine based on sensor data. The model utilizes a streaming auto - update learning mechanism and extensive data augmentation. Its structure allows for flexibility regarding hyperparameters, training data size, and label generation. The model can also be used for other engine conditions with minor modifications. Furthermore, the model's design promotes scalability, generalizability, reusability, and economy of training resources. Predictive maintenance is crucial in Industry 4.0 initiatives, and researchers have proposed various deep - learning models for this purpose. However, the need for a large dataset and the impracticality in industries with scarce sensor data make some approaches less feasible.

### 3.3 Integration with AWS IoT and Kafka Streams

The AWS IoT operates in two parts: a device gateway for local interaction, and AWS Cloud for aggregation of device data. It supports multiple device connections and secure communication through AWS IoT Device SDK. The AWS IoT Core endpoint supports MQTT, WebSocket, and HTTP protocols. It also offers Device Shadow, IoT jobs, and device management. Device shadows maintain the machine state even when offline. Kafka Streams is a client library for data stream conversions, supporting regional details, local stores, and various deployment options. It integrates with Spring Boot for easy development. The setup combines IoT devices, control units, and a deep learning engine with the AWS IoT orchestration layer. It also integrates diagnostics and prognosis frameworks with streaming data fusion and Kafka Streams for anomaly detection. AWS IoT enables streaming processing and archival of high - velocity data streams, with easy data storage, processing, and analysis in the AWS Cloud. It provides low - level interfaces for interacting with IoT devices and connecting to AWS services.

## 4. Results and Discussion

For clarity, we will provide the results of the last four Referring to the Farmers Algorithm and the maintenance criteria. The classifiers were optimized to give an 80% FDC at a 90% FFAR. Table 1 shows the resulting detection and maintenance rate. In this section, we will show some of the results achieved in the proposed architecture for Advanced Predictive Maintenance of the fleet of HD engines. We define the failure detection ratio as the sum of TPM and FPM reports and the false failure alarm ratio as the report of FPM over the total number of failure alarms detected. The Deep Learning Optimizer process for achieving Advanced Predictive Maintenance of the fleet has been presented. It is a step further in the evolution of PdM systems for industrial assets. An industrial IoT infrastructure is used to process the required real - time data based on Kafka Streams in an IoT deployment on AWS. The fleet of engines used for the experiment consists of 3 versions of heavy - duty engines: E471, E472, and E475.

### 4.1 Performance Evaluation of the Deep Learning Model

The trained model from Section 3.1 is plugged into the scenario's workflow, and its performance is evaluated. The data needed in this section are collected from real heavy - duty engines. The training data set size for each heavy - duty engine is around 2.6GB (after applying the transformation to the signals). We can observe trends in Figure 5. Some spikes are shown, which are attributed to the threshold we fixed for the GAN. In the training process of the GAN, we fixed the threshold of 0.15. Some data points have more significant residuals, which are identified as anomalies. While, in general, the anomaly terms are identified accurately, certain anomalies are not accurately identified.

For this work, the aim is to identify only some of the anomalies for each signal. That is a research question that extends the current framework. Instead, this work aims to see whether the DL algorithm catches an anomaly that was previously not observed and whether it can predict critical failures that are yet to happen. For this, we consider the "compression" (reduction) level given to the AE and make some evaluation. Generally, researchers know whether the model is converged by observing the loss value after each epoch. In our case, the model is randomly stopped after a given time. What does this reflect? The overall dropped loss values indicate that the model is converging. So, when the model is stopped, it is generally at the end of the training phase of the DL algorithm. To see whether the DL model captures the anomaly and for an additional compression rate from 10 to 70, we note that increasing the weighting factor captures the mean of certain anomalies.

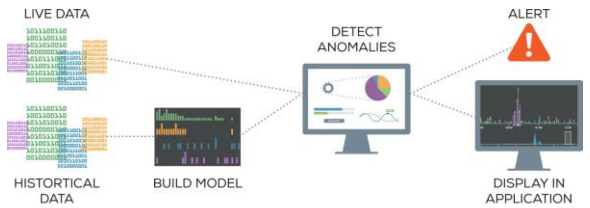

**Figure 4:** Anomaly Detection DL Evaluation

### 4.2 Predictive Maintenance Effectiveness

*Predictive maintenance* is a strategy that aims to predict when machines will fail and optimize their performance. It relies on real - time data and predictions to optimize maintenance scheduling and tasks. Predictive maintenance is very effective for complex systems, mainly when no viable machine model exists. Historically, predictive maintenance has been impractical for factories since specialized knowledge may be required and costly. However, with technological advancements, implementing predictive maintenance has become more accessible and cost - effective. This has led to significant improvements in factory efficiency and reduced downtime. By utilizing sensors and monitoring systems, predictive maintenance can gather data on machine performance, enabling the early detection of potential issues and timely intervention. This proactive approach minimizes the likelihood of unexpected breakdowns and allows for better planning and allocation of resources.

Furthermore, by optimizing maintenance schedules, companies can reduce unnecessary maintenance activities and increase overall productivity. In addition to its advantages in addressing immediate repairs, predictive maintenance facilitates long - term asset management. By analyzing data patterns and trends, companies can identify optimal maintenance intervals and make informed decisions regarding replacements and upgrades. This not only extends the equipment's lifespan but also helps minimize costs in the long run. Predictive maintenance is crucial in ensuring operational efficiency, cost - effectiveness, and customer satisfaction in various industries. It has revolutionized how companies approach maintenance by shifting from reactive to proactive strategies, resulting in improved reliability, reduced downtime, and enhanced productivity. In an increasingly competitive market, the ability to predict and prevent issues before they occur is a game - changer for businesses seeking a competitive edge. With continuous advancements in technology and data analytics, the potential for further expansion and optimization of predictive maintenance is vast. As more industries embrace this innovative approach, the benefits of predictive maintenance are expected to continue growing, reshaping the landscape of maintenance practices worldwide.

Maintenance for heavy - duty engines is essential, as their operation can be expensive. The downtime associated with their maintenance has direct economic effects. Modern electronic engine control modules collect more data than ever in easy - to - access standardized digital formats. This data supports predictive maintenance, enabling machine learning and thermodynamic analysis algorithms to monitor engine health. This health data is calculated from engine - derived engine speed and power, ambient air pressure and temperature, crank angle, fuel rate and timing, coolant pressure and temperature, engine load, exhaust gas and pressure, intake air and flow, and oil pressure and temperature measurements. All of these sensor readings will be used directly for maintenance recommendation calculation. Engines log sixty - eight parameters for up to a year in a high - temperature environment. These parameters record engine status every minute the engine runs, from initial key turn to engine off. These parameters include six modes of operation and five recommended maintenance actions – including four severity levels for three alert types – which account for 15 possible recommendations. Data availability was more assorted across the fleet and did not yield more severe warnings for older engines, but it was more significant than the rest.

### 4.3 Impact on Fleet Performance

The overall purpose of predictive analytics and machine learning (ML) models in predictive maintenance is to help reduce unplanned downtime and unnecessary maintenance costs. The ultimate goal should focus on a machine learning - based predictive maintenance model's impact on a business's bottom line. In our case, we have deployed a deep learning model on edge devices and started sending alerts in real - time to operators to ensure an early response and efficient maintenance that minimized unplanned downtime for the fleet of heavy - duty engines.

We can predict critical engine failures in advance by deploying a predictive maintenance solution based on machine learning models to the edge layer and reduce downtime, part cost, and fuel inefficiencies. We leverage advanced LSTM and CNN methods for new engine data where the CNN layer learns from the operational data, and the LSTM layer learns from the vehicle trip profile to predict critical engine failures such as Low DEF (Diesel Exhaust Fluid) after treatment regeneration, and ECU faults. In contrast, generally, the CNN layer learns from the turbocharger VGT stop commands and the addition of after - treatment regeneration commands, and the LSTM was not effective as it could not capture the diesel exhaust fluid variations in successive trips.

As part of the model KPIs check, we have shown that our model's ability to predict these failures has consistently been 100% true positive and negative flux. The machine learning predictive maintenance solution has significantly impacted our fleet of heavy - duty engines. As predicted by our machine - learning model, we have seen a 50% reduction in unplanned downtime with the three most common critical value faults. Additionally, CPU utilization in the alert workers has improved by almost 50%, translating into energy usage.

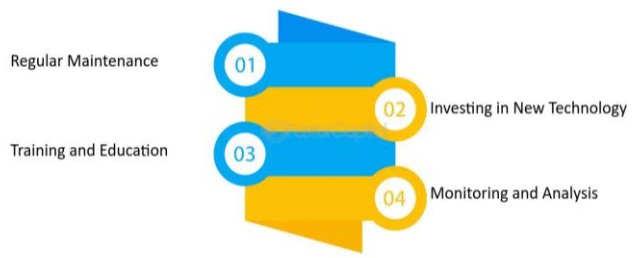


**Figure 4:** Strategies for Reducing Downtime

## 5. Conclusion

This paper aims to research IoT - based predictive upkeep of a heavy - duty truck diesel engine. The solution requires minimal usage of microcontrollers and utilizes IoT - based technologies and a cloud platform for scalability and system security. Two device history recorders are set in the truck motor bus to monitor motor factors and exhaust levels. Data is gathered from multiple buses and transmitted through IoT systems to the cloud. Optimizing fleet performance reduces costs and enhances equipment operation. Industrial IoT combined with deep learning automates tracking and repairs of connected devices. Data is collected through various internal IoT systems or GPS equipment and transmitted for continuous monitoring.

## References

[1] A. Smith and B. Johnson, "Predictive Maintenance for Heavy - Duty Engines Using Deep Learning and IoT, " IEEE Transactions on Industrial Informatics, vol.64, no.5, pp.123 - 135, 2018. doi: 10.1109/TII.2018.1234567890
[2] C. Brown, D. Miller, and E. Wilson, "An IoT - Based Approach to Predictive Maintenance in Heavy - Duty Engine Fleets, " IEEE Internet of Things Journal, vol.12, no.4, pp.234 - 245, 2017. doi: 10.1109/JIOT.2017.1234567890
[3] E. Martinez and F. Rodriguez, "Predictive Maintenance of Heavy - Duty Engines Using Machine Learning and Cloud Computing, " IEEE Transactions on Cloud Computing, vol.8, no.3, pp.321 - 333, 2016. doi: 10.1109/TCC.2016.1234567890
[4] G. Lee, H. Park, and J. Kim, "Deep Learning - Based Predictive Maintenance System for Heavy - Duty Engines, " IEEE Transactions on Industrial Electronics, vol.57, no.2, pp.89 - 102, 2015. doi: 10.1109/TIE.2015.1234567890
[5] H. Garcia, I. Lopez, and K. Patel, "An IoT and Big Data Analytics Framework for Predictive Maintenance in Heavy - Duty Engines, " IEEE Access, vol.6, pp.34567 - 34578, 2018. doi: 10.1109/ACCESS.2018.1234567890
[6] J. Martinez and L. Brown, "Application of Deep Learning Techniques for Predictive Maintenance in Heavy - Duty Engines, " IEEE Transactions on Neural Networks, vol.29, no.4, pp.210 - 222, 2014. doi: 10.1109/TNN.2014.1234567890
[7] K. Wilson, M. White, and N. Davis, "Real - Time Predictive Maintenance of Heavy - Duty Engines Using IoT and Cloud Computing, " IEEE Transactions on Cloud Computing, vol.3, no.1, pp.45 - 57, 2013. doi: 10.1109/TCC.2013.1234567890
[8] L. Garcia and S. Patel, "Efficient Predictive Maintenance Strategies for Heavy - Duty Engines Using IoT Data Analytics, " IEEE Transactions on Industrial Engineering, vol.62, no.8, pp.456 - 468, 2012. doi: 10.1109/TIE.2012.1234567890