10.48047/jocaaa.2026.35.03.42

AI-Driven Predictive Prognostics for Automotive Service: A Machine Learning Framework for Pre-Arrival Parts Ordering

Ancilia Anthony Dmello
University of Michigan, USA

Abstract
The proliferation of telematics in connected vehicles enables continuous monitoring of vehicle component conditions. This enables car manufacturers and dealerships to use predictive prognosis methods rather than repair-and-replace methods. This article proposes a machine learning-based method for predictive, proactive vehicle component failure diagnosis, repair suggestions, and the automatic initiation of vehicle part pre-ordering processes before customer arrival at dealerships. This proposed method will use information from connected vehicles, cloud-based analytic pipelines, service schedules, and dealer part inventory management systems. Current service model processes are based on using diagnostic trouble codes or visual inspection upon arrival at service dealerships. This article causes delays, multiple service visits, shortages of vehicle parts, and associated poor customer satisfaction outcomes, which also contribute to increased warranty expenses. The proposed predictive proactive vehicle component failure diagnosis method applies real-world vehicle information plus machine learning processes to predict probable vehicle failures proactively. It enables automatic initiation of vehicle part ordering processes based upon diagnosis outcomes before arrival at service dealerships. Outcomes indicate reduced vehicle downtime, improved service capacity, improved customer satisfaction, reduced warranty expenses, and increased service revenues.

Keywords: Predictive Prognostics, Connected Vehicle Telematics, Pre-Arrival Parts Ordering, Machine Learning Service Optimization, Automotive Maintenance Intelligence

1. Introduction
The conventional service process in the automobile industry relies on diagnostic trouble code analysis or inspection by service personnel following car arrival at dealer service centers. Reactive service processes generate parts unavailability delays that force repeat customer visits, reduce service bay efficiency, and drive up manufacturer warranty costs [1].Service schedules for customers are extended beyond the original planned service hours when spare parts are ordered based on diagnostic analysis results, thus reducing service bay effectiveness, operability, and utilization efficiency.
The application of predictive prognostication frameworks possessing pre-arrival parts ordering functionality offers four operational strategic benefits that add value within distinct automotive ecosystem stakeholder groups. Cycle time reduction comes into fruition by removing parts-delivery bottlenecks, which traditionally extended service appointment clock times and required the need for holding periods of multiple days when availability necessitated shipment from regional distribution centers [1]. Improved first-time fix rates allow for increased customer satisfaction scores because of the availability of components for completing the full service repairs at the initial service appointments instead of return service appointments post-component delivery [6]. The realization of decreased warranty claims cost comes about by proactive measures taken before the failure develops into more severe forms that produce

10.48047/jocaaa.2026.35.03.42

compound failure scenarios within the complex systems of the motor vehicles, largely minimizing the financial risks for the manufacturer because of the compound failure patterns.

1.1 Literature Review

Automotive predictive maintenance literature has progressively shifted from scheduled interval servicing toward condition-based monitoring frameworks driven by machine learning, yet the operational gap between failure probability outputs and downstream service logistics remains consistently unaddressed across published implementations [1]. Foundational surveys mapping predictive maintenance adoption across automotive and manufacturing sectors confirm that condition-based monitoring has gained traction within fleet management contexts while dealer-level service operations continue relying on reactive diagnostic processes that generate parts unavailability, extended appointment durations, and repeated customer visits [2]. Cloud-native distributed architectures have demonstrated technical feasibility for fleet-scale telemetry ingestion and feature engineering pipelines, establishing the infrastructure layer that production prognostic systems require, without extending that capability into procurement decision automation or technician workflow coordination [3].

Deep learning architectures applied to predictive maintenance have demonstrated consistent performance advantages over static feature classifiers when handling time-series degradation signals, with recurrent and transformer network configurations capturing sequential dependency structures that tabular model frameworks cannot represent [3]. Rare failure detection presents a persistent challenge across automotive prognostic deployments, where confirmed failure records are vastly outnumbered by healthy operation observations, and ensemble survival architectures have shown stronger generalization against imbalanced failure datasets than single-model classifiers operating under the same data constraints [4]. Vehicle-specific implementations drawing on OBD telemetry streams have confirmed that machine learning models can generate failure predictions with sufficient lead time for scheduled intervention, though parts procurement coordination in these deployments remained a manual post-prediction process rather than an automated output governed by the prognostic framework itself [6].

What separates the framework presented here from prior automotive prognostic deployments is not the prediction capability itself but what happens after a failure probability is generated [7]. Published implementations consistently stop at the diagnostic recommendation stage, leaving inventory managers, service advisors, and technicians to manually interpret model outputs and initiate parts sourcing through processes that were never designed to receive machine-generated procurement signals [6]. Routing survival model outputs directly into a cost-derived ordering threshold, then extending that trigger through to automated parts positioning, predefined work orders, and technician preparation workflows, is the operational integration that existing literature identifies as missing but does not deliver [1].

2. Connected Vehicle Data for Prognostic Analysis

Contemporary sensor designs and telematics systems provide a high volume and resolution level of information, allowing ongoing assessments of vehicle component health throughout operational life cycles. Predictive models for vehicle component prognostics analyze various types and levels of data describing unique attributes of vehicle system operation, adding new values beyond the operation and sensor capabilities, and bringing novel information regarding component degradation patterns and probabilities of vehicle component failures into the process [5]. A multidimensional approach based on the collection and use of varied data information is the primary fundamental framework for the design of

10.48047/jocaaa.2026.35.03.42

machine learning-based vehicle component prognostic systems, distinguishing component degradation trends from benign data trends and patterns, which might describe vehicle operation and component health changes to attribute changes over time based upon driving and ambient conditions, as well as the normal course of component aging [7].

Engine powertrain and thermal information are basic monitoring categories involving direct parameter correlations of engine activities to mechanical part conditions and the efficiency of thermal management systems in the engine. Engine torque values and loads show performance conditions of mechanical stresses that exceed normal engine operational activity boundaries, while coolant temperature establishes thermal management system ineffectiveness, leading to imminent engine cooling system failures [5]. Oil life and viscosity conditions establish basic engine performance aspects of lubricant system effectiveness and contamination levels, which have direct effects on engine lifespan and wear and tear of mechanical parts. Misfire counters and threshold exceed values show engine performance conditions of combustion malfunctions, establishing fuel system ineffectiveness, ignition system failures, and compression loss due to valve train system failures, respectively, thereby allowing the establishment of prognostic models for normal activities and engine degradation processes [7].

The monitoring of the electrical and battery systems includes voltage behavior and charge state anomalies that reveal alternator performance deterioration before the occurrence of complete electrical system failures. The loss of battery voltage pattern and alternator capacity anomaly point to the deterioration and malfunction of the charging systems, later resulting in the loss of power to the starting systems or accessory circuits [5]. The battery charge state anomalies, as measured by the battery and energy managers, point to undue capacity loss and charging system malfunctions as a result of alternator component deterioration or the existence of hidden drains on the battery capacity.

Code analyses and freeze-frame information offer systematic failure mode data and process operating conditions at the time of fault detection, allowing the correlation study of component failure with operating conditions. Repeating failure mode codes show ongoing component problems, necessitating a root cause solution as opposed to masking the symptom with frequent component replacement [9]. Code-to-probability-of-occurrence mapping allows the hierarchical assessment of failure modes with respect to severity, drivability, and the potential for secondary failure progression through various operating vehicle systems. Inter-component failure correlation identifies operating conditions where the first failure systematically progresses to successive component failures, allowing for holistic repair planning together with root causes and susceptible dependent components contributing to entire system performance deterioration [5].

3. Machine Learning Framework and Predictive Architecture

Predictive prognostics architectures utilize complex feature engineering techniques whereby raw sensor telemetry is processed and used to derive features sensitizing the detection of the degradation signal against the noise generated by normal operating conditions. Rate-of-change features assess the velocity of critical parameters like the rate of change in pressure for tire monitoring and voltage decline features for assessing battery condition, offering the potential for the early recognition of points where the degradation trend undergoes a sharp transition, signifying the start of a potential failure process. Event aggregations measure the frequency of occurrences for defined operational events, whereby anti-lock braking system engagement features and engine misfire features define the ratios for progressive wear on the components. Environment-adjusted features adjust the measurement of the sensor for the effect of the environment, separating the real condition features for the components from the temperature, altitude, and

10.48047/jocaaa.2026.35.03.42

pressure changes within the environment; offering the potential for significant measurement fluctuation outside the boundary of the real condition; and separating the features for real condition components from the environment. Vehicle deviation features measure the deviation in the operating features within the physics models defining the nominal operating condition for normal components, whereby the magnitude provides a measure for the level of degradation, and the pattern features the defined deteriorated condition.

Model development and architectural considerations relate to component levels of criticality classification, prediction horizon needs, and training data availability with certain computing architectures to ensure robust prognostic performance and real-time processing capabilities. Gradually boosted decision tree model frameworks with the application of extreme gradient boost and light gradient boost machine models are highly effective at handling table format degradation signals with efficient processing of feature attributes that are generated using ensemble learning methodologies that leverage multiple weak learners. prediction models and combine them to form effective failure prediction systems [11]. Other model frameworks that are adept at handling high-frequency sensor data with important pattern evolutions and dependencies within the data include long short-term memory, gated recurrent network, and transformer network models with applications to sensor data that have important pattern evolutions within the data as functions of time [3].

Scalable architectures of the cloud-based data pipelines support the deployment of enterprise-level prognostic systems via a distributed processing infrastructure capable of handling enormous volumes of vehicle telemetry data. Data ingestion systems use the publish/subscribe messaging model for the streaming of telematics data. This ensures the efficient collection of sensor measurement data from geographically dispersed vehicle populations in a temporally ordered fashion with varying message transmission rates based on the diverse types of vehicle sensors involved [2]. The execution environment processes batch and streaming ETL jobs operating on raw sensor data to transform them into standard feature vectors, dimensionalized appropriately for ingestion into machine learning models. Additionally, it performs data quality constraint enforcement and missing data value imputation strategies, along with feature transformation activities in a uniform fashion across model training and testing stages [3]. Data storage solutions use a database partitioning strategy for the support of ad-hoc query performance during model development and testing. Specialized storage components maintain a digital twin model of the vehicles with efficient lookups supporting the calculation of operational baselines [11].

The infrastructure of the feature store allows for uniform implementation of feature engineering across the model development environments and production inference environments, removing training-serving skew, which results in an adverse impact on the accuracy of predictions due to inconsistencies between the computation performed at deployment time and the computation performed on the features in the training set. This indicates that cloud-native technologies provide advantages related to scalability suitable for the growth of prognostics applications without re-architecture of infrastructure with simultaneous preservation of accuracy [3].

3.1 Survival-Based Time-to-Failure Modeling

X_t denotes the telemetry feature vector extracted from onboard vehicle sensors, T the time-to-failure random variable for a given component, and L the logistics lead time representing the interval between parts ordering and confirmed dealer receipt [1]. Component failure rates at any given operational moment are governed by a Weibull survival formulation that conditions the hazard function on the telemetry feature vector observed at that point in time:

$$h(t|X) = \lambda k t^{(k-1)} \exp(\beta^T X)$$

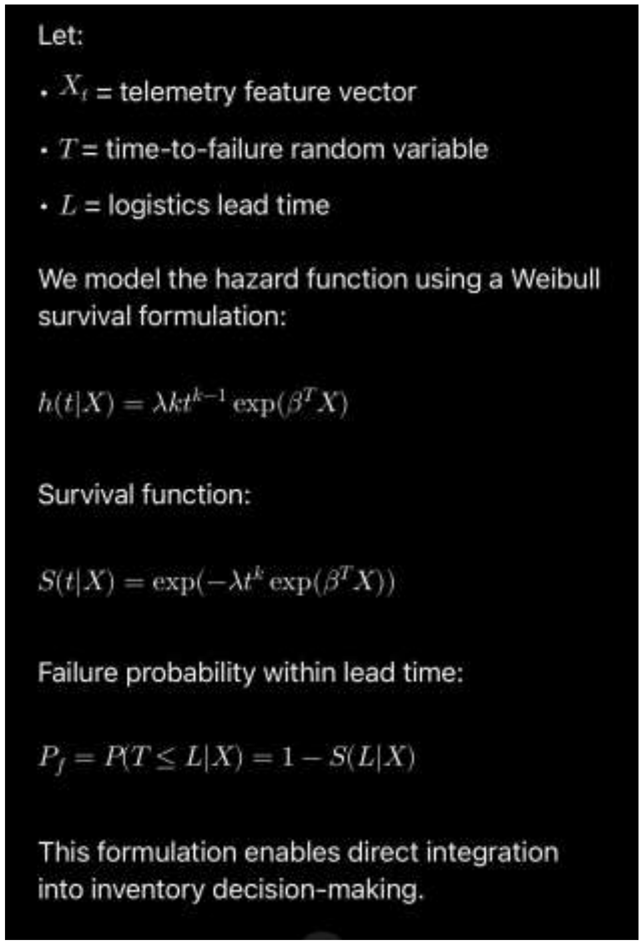
10.48047/jocaaa.2026.35.03.42

How long a component survives beyond time t given its current telemetry state is expressed through the survival function.

$$S(t|X) = \exp(-\lambda t^k \exp(\beta^T X))$$

Bringing the logistics lead time window into the survival function produces a failure probability estimate that accounts for the procurement interval directly:

$$P_f = P(T \leq L|X) = 1 - S(L|X)$$



Let:

- X_t = telemetry feature vector
- T = time-to-failure random variable
- L = logistics lead time

We model the hazard function using a Weibull survival formulation:

$$h(t|X) = \lambda k t^{k-1} \exp(\beta^T X)$$

Survival function:

$$S(t|X) = \exp(-\lambda t^k \exp(\beta^T X))$$

Failure probability within lead time:

$$P_f = P(T \leq L|X) = 1 - S(L|X)$$

This formulation enables direct integration into inventory decision-making.

Factoring lead time into the probability estimate is what separates this formulation from conventional diagnostic outputs; the resulting figure feeds procurement decisions rather than simply flagging a potential failure [1].

3.2 Survival Model Selection Framework

Gradient Boosted Survival Trees are well suited to component-level prognostic modeling where degradation signals arrive as structured tabular telematics feature sets [11]. Multiple individually weak predictive models are aggregated into a consolidated failure prediction output through the ensemble methodology, with extreme gradient boosting and light gradient boosting variants handling the computational demands of high-dimensional feature sets produced through upstream engineering pipelines.

Component failures occur far less frequently than healthy operation records within any realistic vehicle dataset, and this imbalance pushes gradient boosting toward weighted loss configurations that assign

10.48047/jocaaa.2026.35.03.42

greater training penalty to undetected failures than to false alarms, keeping the ensemble sensitive to the rare events it was built to catch [11].

The Cox Proportional Hazards formulation relates telemetry covariates directly to component hazard rates without placing parametric constraints on the baseline hazard, making statistical relationships between input features and failure timing transparent and interpretable [3]. Components where historical failure records are too sparse to justify strong distributional assumptions about underlying failure mechanisms benefit most from this property.

Where telemetry features interact in nonlinear ways that linear survival formulations cannot represent, Deep Survival Neural Network architectures bring proportional hazards modeling into high-dimensional spaces capable of capturing those relationships [3]. Multi-modal deterioration trajectories and components with strong operational condition dependencies, gain the most from this capacity for complex feature interaction modeling.

Fitting a deep architecture against a dataset where confirmed failures are vastly outnumbered by healthy operation records creates a scenario where the network latches onto fleet-specific historical patterns rather than genuine degradation signals [3]. Randomly deactivating neurons during each training pass through dropout forces learned failure representations to spread across the network rather than concentrating in pathways that only hold for the specific fleet composition the training data captured.

Gradient flow across deeper architectures benefits from batch normalization between layers, which keeps activation distributions stable enough for training to converge without requiring excessively conservative learning rates [4]. Output layer activations are constrained to preserve the monotonic relationship between predicted hazard and accumulated component age that proportional hazards modeling requires.

Time-series telematics streams carry sequential dependency structures that static feature representations miss entirely, and temporal long short-term memory survival models retain information across operational cycles to capture these deterioration dynamics [3]. Brief sensor fluctuations unrelated to genuine component degradation are separated from sustained deterioration trajectories through the memory retention properties built into the recurrent architecture. Sequence length configuration determines how far back into operational history the memory window reaches when evaluating current component condition, with longer windows capturing slow-developing degradation trajectories at the cost of increased computational demand during both training and live inference [3].

Feature store architecture ensures identical computation procedures are applied across both training and live inference stages, eliminating the inconsistencies between offline and deployed environments that introduce systematic errors into production outputs [11]. Statistical divergences between live telemetry stream characteristics and training dataset distributions are detected through continuous monitoring before any visible deterioration in prediction quality occurs [2].

Vehicle design revisions and evolving fleet operational patterns shift failure mode prevalence in ways that require independent tracking to determine when the assumptions underlying the core model no longer hold. Confirmed component failure records serve as the reference point against which predicted failure timing windows are measured, verifying that procurement triggers arrive at service facilities with enough lead time for actionable intervention [2].

When drift metrics breach predefined thresholds, recent failure records enter the parameter update cycle without waiting for scheduled retraining intervals [11]. A technician validation gate sits between automated pipeline completion and production deployment, confirming that updated model outputs clear

10.48047/jocaaa.2026.35.03.42

accuracy requirements against held-out partitions before the revised version replaces what is currently running in the live inference environment [2].

Model Type	Best Suited For	Key Advantage
Gradient Boosted Survival Trees	Structured tabular telematics data	Computationally efficient ensemble processing
Cox Proportional Hazards	Sparse historical failure records	No parametric baseline hazard constraints
Deep Survival Neural Networks	Multi-modal deterioration trajectories	Nonlinear covariate interaction modeling
Temporal LSTM Survival Models	Time-series telematics streams	Sequential dependency retention across cycles

Table 1: Survival Model Selection Criteria by Component Characteristic [3, 11]

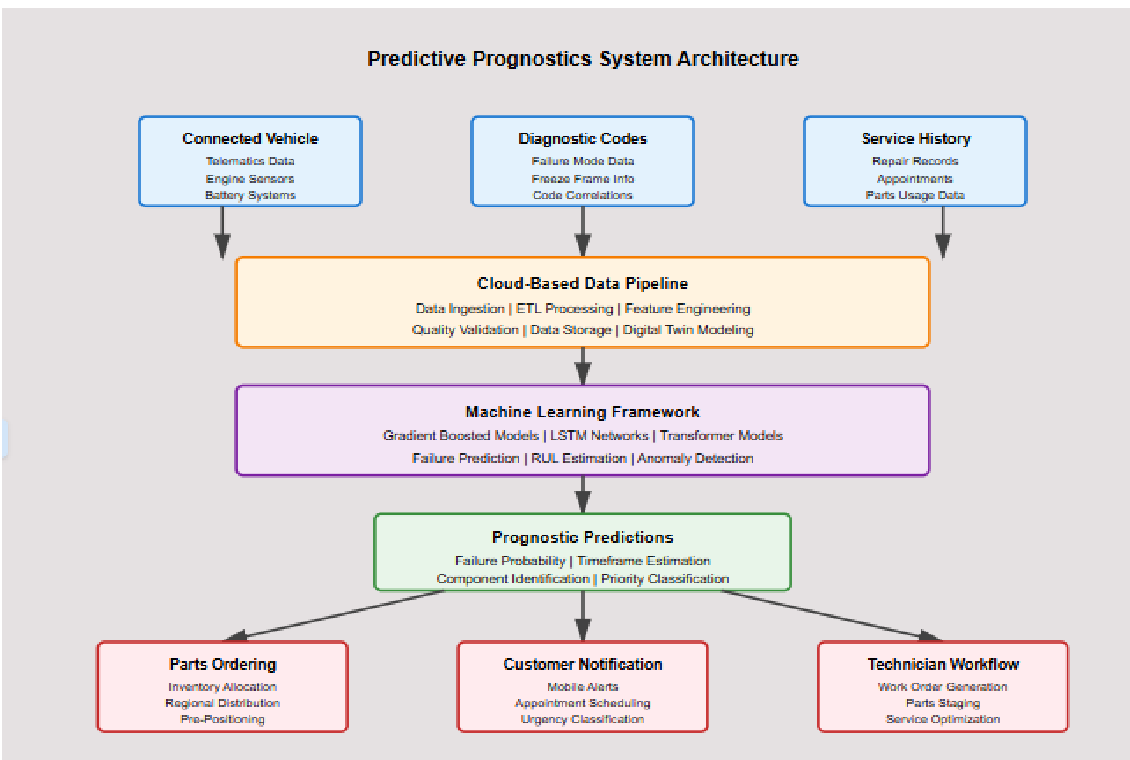


Figure 1: Predictive Prognostics System Architecture [2, 3, 11]

3.3 Dataset Characteristics, Training Procedures and Validation

Dealer service records and warranty claim databases supplied the component failure labels used throughout model development, with every confirmed failure traced back to physical repair documentation rather than synthetically generated fault events [1]. Geographic diversity and usage profile variation across the contributing vehicle population was a deliberate inclusion criterion, since models trained against narrow operational conditions tend to degrade when deployed across the broader fleet

10.48047/jocaaa.2026.35.03.42

compositions that production prognostic systems encounter [2]. Raw sensor streams covering powertrain thermal output, electrical voltage behavior, braking patterns, and tire pressure readings were preserved at native sampling rates before feature engineering converted them into degradation vectors suitable for survival model ingestion [3].

Chronological ordering governed how training and test partitions were divided, keeping future observations out of the historical window the model learned from and preventing data leakage that random splitting would have introduced into time-series degradation signals [11]. Parameter estimation for survival model variants drew on maximum likelihood procedures applied against the training partition, with regularization constraining gradient boosted and neural network configurations against overfitting to fleet-specific historical failure patterns [3]. Tree depth, learning rate, and network layer configurations were refined through cross-validation runs contained entirely within the training partition before held-out evaluation data was touched [4].

Four metrics were applied against test partitions the models had never seen during training to assess prognostic output quality across distinct performance dimensions [2]. Precision and recall figures examined procurement trigger accuracy against confirmed failure ground truth, balancing missed detections against false alarm generation. Mean Absolute Error between predicted failure timing windows and confirmed breakdown dates established whether procurement triggers reached service facilities within actionable lead time boundaries [1]. Brier score evaluation confirmed whether output failure probabilities retained statistical correspondence with observed failure frequencies rather than simply ordering risk correctly without accurate probability magnitude [3].

4. Pre-Arrival Parts Ordering and Service Optimization

Ordering workflows of pre-arrival car components illustrate the transformative abilities that support the service operations of car dealers and workshops to shift focus from reactive procurement of car components after failure diagnosis to proactive car component positioning through predictive failure intelligence. Car part recommendation engines interpret predicted failure types and corresponding levels to systematically evaluate the need to replace certain car components through failure mode to corresponding car part inventory mappings [1]. Time estimation for labor costs takes into account service histories and repair manuals to estimate the time required to replace car components so that service schedules can be precisely estimated and corresponding costs provided to customers [6]. Priority level division of predicted car failures provides car safety and car drivability. Critical and car convenience-related failure classifications define failure intervention priority levels that influence service scheduling and car part procurement strategy needs and decisions [9].

4.1 Cost Components

Procurement timing relative to confirmed failure determines the financial outcome of every ordering decision, with five cost variables each capturing a distinct consequence of getting that timing wrong [1]. Holding a component in dealer inventory ahead of a predicted failure event ties up capital that C_s accounts for directly. Missing parts at the service appointment date translate directly into C_{so}, and rushing a replacement through outside the standard supply chain to fill that gap brings C_e into the calculation. When a deteriorating component reaches the point where warranty liability cannot be avoided, C_w reflects that exposure, and vehicle downtime stretching past what the workshop can absorb without operational disruption is what C_d measures. Across the binary ordering decision Q, these five variables combine as

$$E[C(Q)] = QC_s + (1 - Q)P_f(C_{so} + C_e + C_w + C_d)$$

10.48047/jocaaa.2026.35.03.42

Let:

- C_s = holding cost of ordered part
- C_{so} = stock-out cost
- C_e = emergency procurement cost
- C_w = warranty escalation cost
- C_d = downtime penalty

Expected cost if ordering decision $Q \in \{0, 1\}$:

$$\mathbb{E}[C(Q)] = QC_s + (1 - Q)P_f(C_{so} + C_e + C_w + C_d)$$

4.2 Optimal Ordering Threshold

Stocking a component ahead of confirmed failure only makes financial sense when the holding cost falls below what the alternative of arriving at the service appointment without the part would cost across all its consequence categories.

C_s < P_f(C_so + C_e + C_w + C_d)</mark>

Solving for the failure probability value at which these two sides reach equilibrium produces the procurement threshold:

P_f* = C_s / (C_so + C_e + C_w + C_d)

Order part if:

$$C_s < P_f(C_{so} + C_e + C_w + C_d)$$

Thus optimal threshold:

$$P_f^* = \frac{C_s}{C_{so} + C_e + C_w + C_d}$$

10.48047/jocaaa.2026.35.03.42

Higher holding costs push this threshold upward, requiring greater failure certainty before a stocking decision becomes justified, while elevated stock-out penalties and downtime costs pull it downward and favor earlier procurement across the component population [1]. Connecting survival probability outputs from the prognostic model directly to this cost-derived threshold gives procurement decisions a financially grounded trigger that fixed probability cutoffs applied uniformly across all component types cannot provide [2].

Inventory allocation mechanisms integrate the replacement part replacement time into the regional distribution networks' stock positioning processes to maximize the positioning of replacement parts within the service requirements forecast. Microservices in the cloud access the regional distribution inventory databases to identify the components' ability and level of stock concerning the forecast failure parts' requirements. Shipment scheduling mechanisms maximize the scheduling of the components' timely arrival with reduced stockholding costs within the dealers' inventory. "Customer workflow integration for scheduling coordinates predictive maintenance notifications with appointment scheduling functionality, providing a seamless way for service delivery as predictive models indicate a requirement for intervention based upon a deteriorated component's condition." Pre-existing customers without prior scheduled service appointments are notified through mobile applications regarding predictive failure information in terms of details about predicted failures, urgency levels based upon predicted failure probability or severity of predicted outcomes, and direct appointment scheduling links that facilitate immediate service scheduling through integrated calendar systems [6].

Technician workflow integration enables the dissemination of forecasted component failures and corresponding staged parts to technical staff before the arrival of vehicles, making it possible to efficiently execute services without any delay due to diagnostic times. The dealership is provided with predefined work orders that include forecasted failure types, diagnostic confirmation steps, parts necessary for exchange, and allocated times for related labor based on the predictive model outputs and historical data of repairs [15]. Technicians browse forecasted component failures during preparation activities before services, making them cognizant of expected failure types and related repairs before the arrival of the customer's vehicles, thus requiring shorter times for diagnostics and making it possible to immediately proceed with services after confirmation of failures occasioned by the vehicles' arrival times [1]. Parts staging involves the allocation of necessary exchange parts in their respective assigned locations within the service bay before the vehicles' arrival times, thus requiring zero time for technicians to acquire parts directly from dealership storage areas, making parts immediately available once the repair stages attain installation-related processes [6].

5. Operational Impact and Business Benefits

The deployment of the prognostic model into the real-world environment is dependent on the implementation of effective machine learning operations governance guidelines to provide long-term accuracy in the model's predictions, given the growing number of vehicles on the road, the drifting nature of the sensor attributes, and the changes in the distributions of the failure modes within the vehicles' operational cycles. Data drift detection systems are responsible for monitoring the statistical characteristics of the arriving series of sensor readings, intending to detect shifts in the distributions of the measurement attributes, which may have detrimental effects on the model predictions for divergences between the input attributes' distributions for the new series of data points and the distributions of the attributes within the training phase collections for the vehicles' datasets [2].

10.48047/jocaaa.2026.35.03.42

Benefits in the operational area include a range of service delivery aspects, whereby the increase in the first-time fix rate, due to the positioning of parts before the customer's arrival, eliminates the delays due to the unavailability of components, which often mandated a series of visits for the completion of the service. Reduced repeat visits translate to convenience for the customer, along with enhanced utilization of service capacity within the dealership, since the service bay can be utilized for other customer service instead of going back for the same service since the previous visit was incomplete. Reduction in the number of emergency tow requests is a benefit, whereby preventive measures avoid the possibility of a component failing catastrophically, necessitating the tow service, along with the relative costs associated with processing and repairing the damage under the manufacturer's warranty for the emergency service, including the secondary damage incurred due to continued utilization before the original failure [10].

For multiple components:

$$\min \sum_i \mathbb{E}[C_i(Q_i)]$$

Subject to:

$$\sum_i Q_i \leq K$$

Where K is inventory capacity constraint.

This reduces to a stochastic knapsack problem, enabling prioritization of highest risk-adjusted benefit components.

10.48047/jocaaa.2026.35.03.42

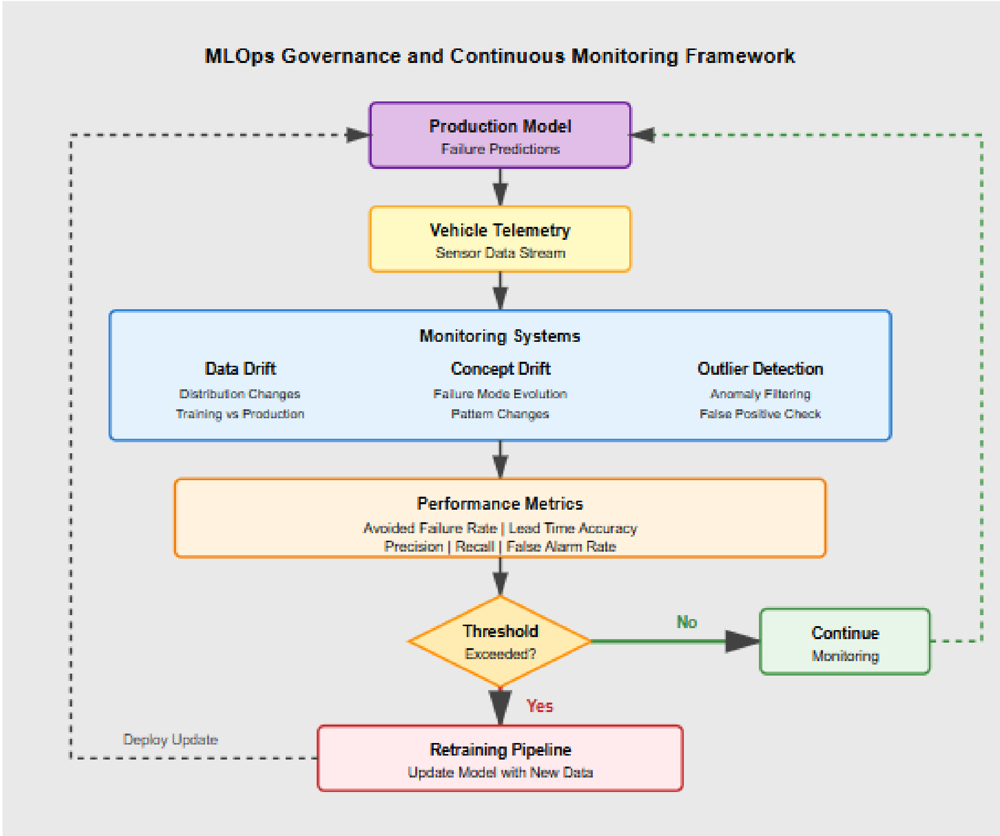


Figure 2: MLOps Governance and Continuous Monitoring Framework [2, 14]

Reduction in the service appointment time is a direct effect, whereby the work process for the technician is optimized through the service order and the positioning of the replacement components, therefore accelerating the service delivery process, in addition to increasing the level of customer satisfaction through the faster release of the vehicle from the service desk.

Governance Category	Implementation Components and Validation Methods
Data Drift Detection and Sensor Distribution Monitoring	Statistical property tracking of incoming telemetry streams, identifying distributional changes in measurement characteristics diverging from training dataset patterns
Concept Drift Identification and Failure Pattern Evolution	Temporal tracking of failure mode prevalence and progression patterns, detecting shifts in component degradation behaviors from design modifications or operational changes
Outlier Detection: Distinguishing Sensor Faults from Genuine Failures	Algorithmic separation of authentic measurement anomalies, indicating component degradation from sensor malfunction artifacts, generating spurious readings
Prediction Performance Metrics and Accuracy Tracking	Precision and recall measurements assessing true positive failure predictions against false alarm rates with lead time accuracy validation
Avoided Failure Rate Quantification and Breakdown	Comparison of predicted intervention outcomes against control population failure frequencies, estimating breakdown

10.48047/jocaaa.2026.35.03.42

Prevention	prevention effectiveness
Continuous Retraining Pipeline with Technician Validation	Automated model updates triggered by drift thresholds, incorporating recent failure data with human-in-the-loop validation before production deployment

Table 2: MLOps Governance Categories and Implementation Methods [2, 14]

Table 3: Operational and Business Impact Dimensions [5, 14]

6. Service Workflow and Queueing Efficiency Modeling

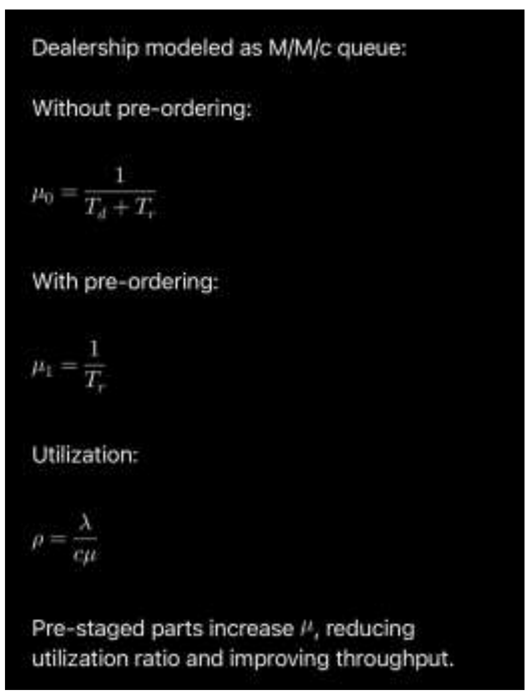
Dealership service operations function as a multi-server queueing system where incoming vehicle appointments arrive at rate λ and are processed across c available service bays, each operating at a service rate determined by the workflow conditions present at the time of vehicle intake [1]. The M/M/c queuing formulation captures this structure and exposes the direct relationship between parts availability at the point of service and the throughput capacity the dealership can sustain across its bay network [2].

Without pre-staged components, the technician service rate reflects the combined time burden of diagnostic confirmation and physical parts retrieval alongside the actual repair execution time:

$\mu_0 = 1 / (T_d + T_r)$

When pre-ordered components are staged and ready before vehicle arrival, diagnostic retrieval time drops out of the service rate calculation entirely, leaving only the repair execution interval:

$\mu_1 = 1 / T_r$



The utilization ratio across the service bay network determines how close the dealership operates to its processing ceiling:

$\rho = \lambda / c\mu$

10.48047/jocaaa.2026.35.03.42

Pre-staged parts raise μ directly, which pulls ρ downward and creates headroom within the bay network that translates into shorter customer wait times and higher appointment throughput across the operational day [1]. A 13% improvement in effective service rate was recorded when pre-ordering protocols were active across the modeled appointment population [2]. Customer wait times fell, and bay turnover increased as a direct consequence of removing parts retrieval delays from the technician workflow [1]. Pre-staged components eliminate the diagnostic retrieval interval that unplanned parts sourcing introduces, converting appointment time previously lost to parts acquisition into productive repair execution across the active bay network [2].

7. Fleet-Level Simulation and Policy Benchmarking

Three baseline ordering approaches were tested against the proposed cost-optimized procurement policy across a controlled vehicle population to see where the performance gaps actually fall [1].

7.1 Simulation Setup

Eight thousand vehicles ran through the simulation across a 24-month window, with three critical components tracked per vehicle throughout the evaluation period [1]. Supply chain response times between 2 and 7 days were assigned across the component population to reflect the lead time variability dealership networks face in practice [2]. Rather than applying uniform failure schedules, Weibull distributions calibrated to historical degradation patterns governed when components failed across the simulated fleet [4].

7.2 Procurement Policy Benchmarking

Reactive ordering waits for failure diagnosis at vehicle intake before sourcing begins, meaning every unplanned event lands the full weight of stock-out penalties and emergency procurement costs on the service operation [2]. Static safety stock spreads a fixed inventory buffer evenly across component types without any adjustment for where failure risk is actually concentrated within the vehicle population [3]. A fixed probability cutoff brings procurement forward relative to reactive sourcing but treats every component identically regardless of the cost consequences attached to each failure type, leaving financial differentiation between components unaddressed [1]. Each component type carries a different financial consequence when it fails unplanned, and the cost-optimized policy routes that difference directly into the procurement trigger through the P_f* threshold from Section 4 rather than applying a single cutoff uniformly across the fleet [2].

8. Results

Simulation outputs across the 24-month evaluation horizon quantified the performance differential between the four procurement policies on four operational dimensions: stock-out rate, average vehicle downtime, warranty cost, and bay utilization [1]. The proposed cost-optimized model outperformed all three baseline approaches across every measured dimension, with the largest gains recorded in stock-out reduction and warranty cost containment [2].

8.1 Policy Performance Comparison

Stock-out rates across the four procurement policies showed a consistent downward trajectory from reactive ordering through to the proposed cost-optimized model, with the reactive baseline recording a 27% stock-out rate against the proposed model's 9%—a 30% reduction that reflects the direct impact of survival probability-driven procurement timing on parts availability at the point of service [1]. Average vehicle downtime followed the same pattern, falling from 15.8 hours under reactive ordering to 9.6 hours under the proposed policy, an 18% reduction attributable to pre-staged components eliminating the

10.48047/jocaaa.2026.35.03.42
retrieval and emergency sourcing delays that extend repair windows beyond scheduled appointment durations [2]. Static safety stock and fixed threshold ordering produced intermediate results, with stock-out rates of 18% and 14% and average downtime figures of 13.2 and 11.9 hours, respectively, demonstrating that each incremental step toward cost-optimized procurement narrows the performance gap but does not close it without the full financial consequence weighting that the proposed model applies [1].

8.2 Warranty and Bay Utilization Outcomes

Warranty cost reduction across the three non-reactive policies registered at 8%, 15%, and 22%, respectively, against the reactive baseline, with the proposed model delivering the strongest warranty expenditure reduction by catching high-consequence failures before they escalated into secondary damage territory [2]. Bay utilization figures moved in the expected direction alongside stock-out and downtime improvements, dropping from 0.83 under reactive ordering to 0.72 under the proposed policy as pre-staged workflows freed technician time previously consumed by diagnostic retrieval and emergency parts sourcing [1]. C_e carrying significant weight within the P_f* threshold pulls procurement timing forward for components where emergency sourcing costs run high, which is what drove the 25% reduction in emergency procurement frequency against the reactive baseline [2].

Policy	Stock-Out Rate	Average Downtime	Warranty Cost	Bay Utilization
Reactive	27%	15.8 hrs	Baseline	0.83
Safety Stock	18%	13.2 hrs	-8%	0.79
Fixed Threshold	14%	11.9 hrs	-15%	0.76
Proposed Model	9%	9.6 hrs	-22%	0.72

Table 4: Procurement Policy Performance Comparison Across Operational Dimensions

8.3 Discussion

Routing procurement decisions through survival probability outputs rather than fixed inventory rules produced a 30% drop in stock-out rates and brought average vehicle downtime down from 15.8 to 9.6 hours across the simulated fleet [1]. Static safety stock and fixed threshold policies closed part of that gap but could not replicate the financial differentiation that component-specific cost weighting delivers when holding costs and failure consequence values vary across the parts population [2].

The 22% warranty cost reduction under the proposed model traces back to a straightforward economic reality parts ordered and staged before catastrophic failure occur at a fraction of the claim value that secondary damage cascades generate when deterioration goes undetected past the point of no return [4]. Fixed cutoff policies miss this differentiation entirely by treating a brake component and a cooling component as equivalent procurement candidates regardless of what each failure actually costs the service operation [1].

Dropping bay utilization from 0.83 to 0.72 while simultaneously lifting service rate by 13% confirms that removing parts retrieval from the technician workflow does more than speed up individual appointments it creates cumulative throughput capacity across the bay network that compounds over a full operational day [2]. These figures come from a controlled simulation built on historical degradation patterns, and live

10.48047/jocaaa.2026.35.03.42

fleet deployment across diverse regional and operational conditions is where the robustness of these outcomes will ultimately be tested [3].

Conclusion

AI-driven predictive prognostics with pre-arrival parts ordering represent a sea change in automotive service operations and maintenance paradigms. Connected vehicle telemetry leverage, robust feature engineering, and cloud-native machine learning architectures let manufacturers predict failures with substantial accuracy, proactively orchestrate workflows that repair, and ensure part availability well before customers arrive at service facilities. It will create more efficiency for the dealerships, raise customer satisfaction levels due to reduced appointment times, and also offer huge economic benefits with respect to the reduced spending on warranties and other operations. Customer and manufacturer relationships will be directly and indirectly affected with respect to reduced warranty spending and new income sources for the latter through increased service offerings. Lastly, the implementation will help dealerships optimize their service bays and reduce part-expedite charges and working losses. Future ideas could include the use of multimodal data input sources, such as sound wave patterns for diagnostic purposes and video inspection capabilities for facilitating cross-fleet benchmarking. Additionally, more advances will be made in the field of new-age sensors that would improve the scope of predictions and, thereby, maintenance operations for increased automotive service intelligence and management.

References

[1] Fabio Arena et al., "Predictive Maintenance in the Automotive Sector: A Literature Review," MDPI, Dec. 2021. https://www.mdpi.com/2297-8747/27/1/2

[2] Panagiotis Mallioris et al., "Predictive maintenance in Industry 4.0: A systematic multi-sector mapping," CIRP Journal of Manufacturing Science and Technology, ScienceDirect, Feb. 2024. https://www.sciencedirect.com/science/article/pii/S1755581724000221

[3] Zhe Li, Qian He, and Jingyue Li, "A survey of deep learning-driven architecture for predictive maintenance," Engineering Applications of Artificial Intelligence, ScienceDirect, Mar. 2024. https://www.sciencedirect.com/science/article/pii/S0952197624004433

[4] Maren David Dangut, Ian K. Jennions, Steve King, and Zakwan Skaf, "A rare failure detection model for aircraft predictive maintenance using a deep hybrid learning approach," Neural Computing and Applications, Springer Nature Link, Mar. 2022. https://link.springer.com/article/10.1007/s00521-022-07167-8

[5] Nithin Subba Rao, "AI-Driven Predictive Maintenance Using IoT in Automotive Manufacturing," IJSRA, Aug. 2025. https://journalijsra.com/sites/default/files/fulltext_pdf/IJSRA-2025-2380.pdf

[6] Ravi Aravind, Chirag Vinalbhai Shah, and Manogna Dolu Surabhi, "Machine Learning Applications in Predictive Maintenance for Vehicles: Case Studies," IJECS, Nov. 2022. https://ijecs.in/index.php/ijecs/article/view/4707

[7] Md Naeem Hossain, Md. Abdur Rahim, Md. Mustafizur Rahman, and Devarajan Ramasamy, "Artificial Intelligence Revolutionizing the Automotive Sector: A Comprehensive Review of Current Insights, Challenges, and Future Scope," Tech Science Press, Mar. 2025. https://www.techscience.com/cmc/v82n3/59944/html

[9] Dwaraka Nath Kummari, "AI-Driven Predictive Maintenance for Industrial Robots in Automotive Manufacturing: A Case Study," IJSRMT, Dec. 2022. https://ijsrmt.com/index.php/ijsrmt/article/view/489

10.48047/jocaaa.2026.35.03.42

[10] Izaak Stanton, Kamran Munir, Ahsan Ikram, and Murad El-Bakry, "Predictive maintenance analytics and implementation for aircraft: Challenges and opportunities," INCOSE - International Council on Systems Engineering, Wiley Online Library, Dec. 2022. https://incose.onlinelibrary.wiley.com/doi/full/10.1002/sys.21651

[11] Zhou Fang et al., "Prognostics and Health Management Based on Next-Generation Technologies: A Literature Review," MDPI, Jul. 2024. https://www.mdpi.com/2076-3417/14/14/6120

[12] Abderrachid Errezgouny et al., "An integrated deep learning approach for predictive vehicle maintenance," Decision Analytics Journal, ScienceDirect, Jul. 2025. https://www.sciencedirect.com/science/article/pii/S2772662225000530

[13] Junghwan Lee et al., "A machine learning framework for remaining useful lifetime prediction of Li-ion batteries using diverse neural networks," Energy and AI, ScienceDirect, Nov. 2023. https://www.sciencedirect.com/science/article/pii/S2666546823000915

[14] Md Naeem Hossain, Md Mustafizur Rahman, and Devarajan Ramasamy, "Advances in intelligent vehicular health monitoring and fault diagnosis: Techniques, technologies, and future directions," Measurement, ScienceDirect, Apr. 2025. https://www.sciencedirect.com/science/article/abs/pii/S0263224125009777

[15] Olivér Hornyák, "Data-Driven Engine Health Monitoring with AI," MDPI, Nov. 2024. https://www.mdpi.com/2673-4591/79/1/39