

UDC 004.94

DOI [https://doi.org/10.15589/znп2026.1\(503\).2.8](https://doi.org/10.15589/znп2026.1(503).2.8)

COMPARE ONE-STAGE AND TWO-STAGE SEGMENTATION FOR DETECTING VEHICLE DAMAGE IN AN IMAGE

РОЗРОБКА МОДЕЛІ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ ДІАГНОСТИКИ ТРАНСПОРТУ З УРАХУВАННЯМ ВАРТОСТІ ПОМИЛОК

Nataliya I. Boyko

Nataliya.i.boyko@gmail.com

ORCID: 0000-0002-6962-9363

Yurii B. Bodnar

yurii.bodnar.shi.2022@lpnu.ua

ORCID: 0009-0005-7545-5987

Н. І. Бойко,

канд. екон. наук, доцент

Ю. В. Боднар,

здобувач вищої освіти

*Lviv Polytechnic National University, Lviv**Національний університет «Львівська політехніка», м. Львів*

Abstract. Purpose. Development of an intelligent predictive maintenance model for Scania truck Air Pressure Systems (APS) to minimize operational costs under conditions of critical class imbalance and asymmetric error costs.

Method. The methodology is based on a comparative analysis of ensemble machine learning algorithms, including Random Forest, XGBoost, and CatBoost. To address data sparsity, a specific imputation strategy was implemented where missing values are treated as informative features. The research utilizes cost-sensitive learning techniques to adapt loss functions to an asymmetric 1:50 cost ratio (false positives vs. false negatives). Furthermore, decision threshold optimization (Threshold Tuning) was applied to shift the model's classification boundary in favor of higher recall for the minority "failure" class.

Results. Experimental results demonstrate that the CatBoost algorithm provides the highest economic efficiency, achieving a total operating cost of 8,640 units, which is 12.7% lower than XGBoost and 20.8% lower than Random Forest. CatBoost correctly identified almost all critical failures, resulting in only 4 false negatives compared to 9 for XGBoost. A modular software prototype was developed using the FastAPI framework, demonstrating the feasibility of deploying the optimized model into real-time diagnostic environments with low latency.

Scientific novelty. The advantage of the CatBoost architecture for industrial data with high imbalance is justified, and a method for direct integration of a financial cost function into the decision threshold selection process is proposed.

Practical importance. The practical value lies in the developed software complex and the optimized diagnostic strategy, which allows logistics companies to reduce unplanned downtime and emergency repair costs by an estimated 10–15%. The modular architecture of the prototype enables rapid integration into existing fleet management systems. The results provide a ready-to-use framework for engineers and data scientists to build robust predictive maintenance systems for Scania heavy-duty trucks and similar industrial equipment.

Key words: predictive maintenance; machine learning; Scania APS; CatBoost; XGBoost; class imbalance; cost minimization.

Анотація. Мета. Розробка інтелектуальної моделі предиктивного обслуговування пневматичної системи (APS) вантажівок Scania для мінімізації операційних витрат в умовах критичного дисбалансу класів та асиметричної вартості помилок.

Методика. Методологія базується на порівняльному аналізі ансамблевих алгоритмів машинного навчання: Random Forest, XGBoost та CatBoost. Для вирішення проблеми розрідженості даних впроваджено стратегію імпутації, де пропущені значення розглядаються як інформативні ознаки. У роботі використано методи вартісно-орієнтованого навчання для адаптації функцій втрат до асиметричного співвідношення витрат 1:50. Також застосовано оптимізацію порогу прийняття рішень (Threshold Tuning) для зміщення межі класифікації на користь максимізації повноти (Recall) для міноритарного класу.

Результати. Експериментальні результати демонструють, що алгоритм CatBoost забезпечує найвищу економічну ефективність, досягаючи показника загальних витрат у 8 640 од., що на 12,7% нижче за XGBoost і на 20,8% за Random Forest. CatBoost ідентифікував майже всі критичні несправності, допустивши лише 4 про-

пуски проти 9 у XGBoost. Розроблено модульний програмний прототип на базі фреймворку FastAPI, що підтверджує можливість інтеграції оптимізованої моделі в діагностичні середовища реального часу з низькою затримкою.

Наукова новизна. Обґрунтовано перевагу архітектури CatBoost для промислових даних із високим дисбалансом та запропоновано метод прямої інтеграції фінансової функції вартості в процес вибору порогу прийняття рішень.

Практична значимість. Практична цінність полягає у розробленому програмному комплексі та стратегії діагностики, що дозволяє логістичним компаніям знизити витрати на аварійний ремонт та простої на 10–15%. Модульна архітектура прототипу забезпечує швидку інтеграцію в існуючі системи управління автопарком. Результати надають готову методичну базу для побудови надійних систем предиктивного обслуговування вантажівок Scania та аналогічного промислового обладнання.

Ключові слова: предиктивне обслуговування; машинне навчання; Scania APS; CatBoost; XGBoost; дисбаланс класів; мінімізація витрат.

PROBLEM STATEMENT

Previously, in [1], the problem of the increasing complexity of modern heavy-duty vehicles and the need for rapid processing of diagnostic information to ensure the reliability of logistics operations were addressed. As the number of sensors in vehicles grows, traditional methods of scheduled maintenance are becoming economically inefficient. This article establishes a concept for transitioning from reactive repairs to predictive maintenance.

The automation of the monitoring process described in [1] is a critical stage, where the primary challenge is not merely detecting a malfunction but proactively forecasting component failure. To achieve this, machine learning and data mining techniques must be applied – specifically, investigating anomalies in the telemetric sensor readings of the Air Pressure System (APS).

In this study, we analyze classification methods under conditions of imbalanced data and select the optimal algorithm based on the criterion of financial loss minimization (cost-sensitive learning), comparing the effectiveness of ensemble methods and gradient boosting.

ANALYSIS OF RECENT RESEARCH AND PUBLICATIONS

The relevance of predictive maintenance (PdM) has become a subject of intense discussion in the scientific community, as the integration of Industry 4.0 concepts into the logistics sector necessitates a transition from scheduled preventive maintenance to condition-based maintenance.

One of the foundational works in this field is the study [2], in which the authors were among the first to address the problem of using logged vehicle data to predict compressor failures. They demonstrated that combining maintenance records with onboard telemetry data enables the identification of equipment degradation patterns well before actual failures occur, forming the basis for reliable predictive models.

The specificity of diagnosing the air pressure system (APS) in Scania heavy-duty vehicles was formalized within the framework of a challenge described in [3]. The authors of this work detailed the structure of sensor-generated data and identified the main challenges, including high dimensionality, noise, and severe class imbalance.

Notably, this publication was the first to propose framing the problem not merely as a classification task, but as the minimization of economic risk.

A substantial body of recent research focuses on the selection of optimal machine learning algorithms. In particular, study [4] presents a comparative analysis of supervised machine learning methods for technical diagnostics. The authors emphasize that for complex technical systems, such as air conditioning or air pressure systems, classical linear models are ineffective, whereas ensemble-based methods exhibit significantly higher robustness to noisy data.

This argument is further developed in [5], where machine learning techniques were applied to categorize failures specifically within the APS of Scania trucks. The experimental results confirmed that tree-based methods, including Random Forest and gradient boosting algorithms, handle dataset-specific attributes – such as histograms and event counters – more effectively than neural networks, which typically require substantially larger training datasets.

A separate and critically important research direction is addressed in [6], which focuses on cost-sensitive learning. The authors convincingly demonstrate that standard performance metrics, such as accuracy, are misleading when dealing with rare events such as failures. In logistics applications, the cost of a missed failure (false negative) may exceed the cost of an unnecessary inspection (false positive) by an order of magnitude. Consequently, the key objective shifts from maximizing classification accuracy to optimizing the decision threshold in accordance with a business-specific cost matrix.

An analysis of the reviewed literature indicates that while solutions based on Random Forest and XGBoost have been extensively explored, the potential of the CatBoost algorithm combined with direct optimization of a financial metric (total cost) for Scania-related predictive maintenance tasks remains insufficiently investigated, which defines the objective of this study.

SEPARATION OF PREVIOUSLY UNRESOLVED PARTS OF THE OVERALL PROBLEM

An analysis of the reviewed literature indicates that while solutions based on Random Forest and XGBoost

have been extensively explored in the context of technical diagnostics, the potential of the CatBoost algorithm combined with direct optimization of a financial metric (total cost) for Scania-related predictive maintenance tasks remains insufficiently investigated. Specifically, there is a need for a deeper analysis of the CatBoost architecture's effectiveness under the specific data imbalance of the APS system, where the cost of a missed failure is fifty times higher than the cost of a false alarm. This gap defines the objective and necessity of the current study.

PURPOSE OF THE STUDY

The purpose of this study is to develop an intelligent solution for predicting Air Pressure System (APS) failures in Scania trucks. The work aims to analyze operational telemetry data to facilitate the transition from reactive repair methods to a Predictive Maintenance model.

A key aspect of this objective is to minimize the Total Cost of operation by optimizing the model in response to asymmetric financial risks. The research focuses on selecting the most effective machine learning algorithm capable of performing reliably under conditions of significant class imbalance and high data sparsity.

METHODS, OBJECT, SUBJECT AND METHODS OF RESEARCH

The object of the study is the processes of technical diagnostics and predictive maintenance of heavy-duty vehicles based on the analysis of large-scale telemetric datasets.

The subject of the study is machine learning models and algorithms for classifying Air Pressure System (APS) states, operating under conditions of critical class imbalance and asymmetric error costs.

Research methods are based on the following:

- Data Mining: preprocessing sparse data and treating missing values as an informative category.
- Ensemble Learning: application of Random Forest, XGBoost, and CatBoost (utilizing symmetric trees and Ordered Boosting).
- Cost-Sensitive Learning: adapting loss functions to a 1:50 error cost ratio.
- Threshold Tuning: optimizing decision boundaries to minimize financial risks (Type II errors).
- Software Engineering: developing a diagnostic prototype using the FastAPI framework.

THE MAIN MATERIAL

Evolution of maintenance strategies: from reactive to predictive. In modern reliability theory and the operation of complex technical systems, including Scania heavy-duty vehicles, the choice of a maintenance paradigm critically affects the cost-effectiveness of logistics operations. Maintenance strategies are commonly classified as reactive, preventive, or predictive [7], each reflecting a different balance between monitoring costs and downtime losses.

Reactive maintenance (run-to-failure) minimizes diagnostic expenses but is inefficient for

commercial transport due to unpredictable failures that cause unplanned downtime and costly emergency repairs, corresponding to high-impact Type II errors (false negatives). Preventive maintenance relies on scheduled interventions based on MTBF; however, by ignoring the actual condition of components, it often leads to premature replacements and increased costs, equivalent to Type I errors (false positives).

Predictive maintenance (PdM), enabled by continuous telemetry-based condition monitoring, represents the most advanced approach within Industry 4.0. By forecasting failures and identifying the optimal time for intervention, PdM reduces breakdown risk while maximizing component lifetime, as supported by the cost-curve analysis shown in Figure 1.

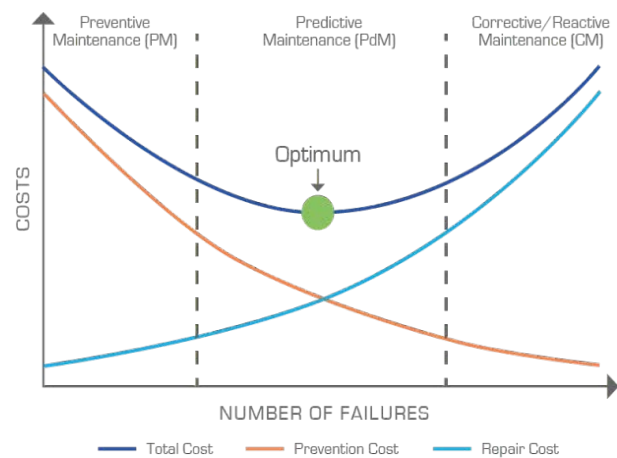


Fig. 1. Comparison of the economic efficiency of different maintenance strategies

Source: created by [8]

As illustrated in the Figure 1, the total cost function is defined as the sum of prevention costs and repair costs. Under a reactive strategy, repair costs dominate and increase due to the unpredictability of failures, whereas intensive preventive maintenance leads to a linear growth in prevention costs. The global minimum of the total cost function is achieved under a predictive approach, which balances failure risk and maintenance expenses. For Scania trucks, where the cost of a missed failure is fifty times higher than the cost of an inspection, adopting a predictive model is a necessary condition for minimizing financial risks.

Monitoring and telemetry data collection technologies. Historically, the earliest and most widely used approach to diagnosing industrial systems is the rule-based method employing fixed threshold values. The essence of this method is to continuously compare the current sensor readings with predefined "normal" ranges. If the value of the parameter X goes outside the interval $[X_{min}, X_{max}]$, the system generates an alarm (Alert) [7]. Despite its simplicity and interpretability, this method has

significant limitations for predictive analytics in complex heavy-duty vehicle systems.

First, threshold-based diagnostics are inherently univariate and unable to capture anomalous correlations among multiple parameters, which are typical of complex mechatronic systems such as the air pressure system (APS). Second, static thresholds ignore operational context, leading to a high rate of false positives. Third, detection is reactive in nature, as threshold violations occur only when degradation has reached a critical stage, thereby confirming failures rather than predicting them. These limitations necessitate a shift toward intelligent data-driven methods capable of identifying hidden patterns in high-dimensional feature spaces.

The effectiveness of predictive maintenance strongly depends on the timeliness and completeness of telemetry data. In modern fleet management systems, two fundamental data transmission paradigms are employed: batch processing and real-time streaming. Batch processing involves collecting telemetry data onboard the vehicle and periodically transmitting it to a centralized server in large packets. This approach reduces communication costs and enables computationally intensive offline analysis of historical data. However, it introduces substantial delays and prevents immediate detection of sudden APS failures.

In contrast, real-time streaming relies on continuous transmission of critical sensor data as it is generated,

enabling near-instantaneous detection of hazardous conditions. While this approach minimizes response latency, it requires a persistent and reliable mobile connection, significantly increasing operational costs, particularly in international freight operations.

An analysis of both paradigms (Fig. 2) shows that neither approach alone is optimal for real-world Scania truck operations. Therefore, this study motivates the development of a hybrid data acquisition architecture (Fig. 3) that combines real-time transmission of critical alerts with batch-based transfer of complete historical data. This hybrid strategy reduces communication costs while maintaining operational awareness and supporting long-term predictive modeling of component degradation.

Characteristics of the Diagnostic Object and Input Data Structure. The air pressure system (APS) is one of the most critical subsystems of modern heavy-duty vehicles, supplying compressed air to braking systems, pneumatic suspension, gear shifting mechanisms, and auxiliary components. As noted in [2], an APS failure during operation triggers emergency braking and results in a complete vehicle stop, causing substantial financial losses due to delivery delays and vehicle recovery.

The core component of the APS is an engine-driven air compressor that compresses ambient air and delivers it to the air processing unit (APU). During air preparation, moisture is removed by a desiccant filter, as water

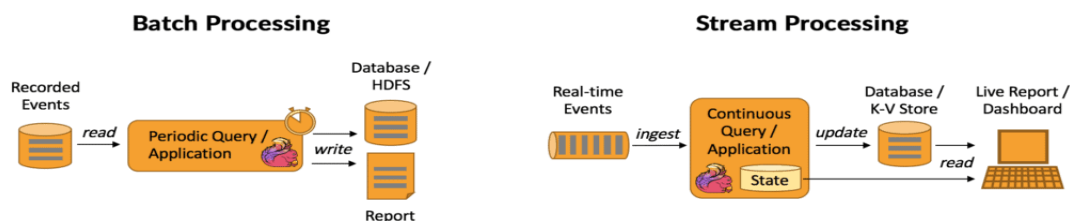


Fig. 2. Comparative diagram of batch and real-time data processing architectures
Source: created by [9].

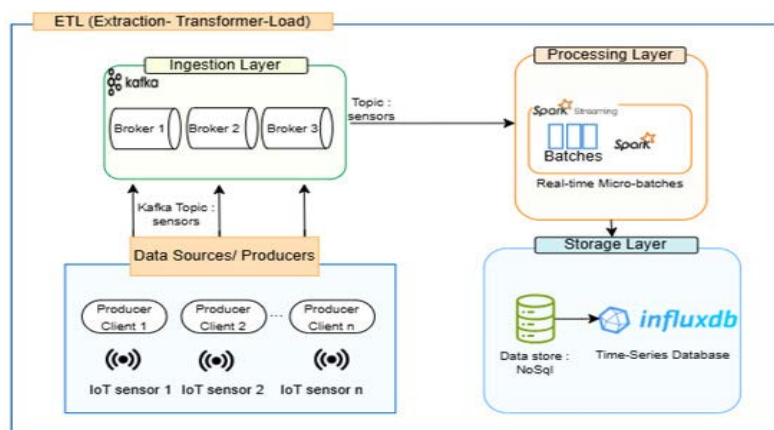


Fig. 3. Scheme of the proposed hybrid data acquisition architecture
Source: created by [10].

in pneumatic lines may cause corrosion and valve freezing, particularly in winter conditions [3]. The dried air is then distributed through a multi-circuit protection valve to separate reservoirs supplying independent subsystems.

Operational reliability analysis identifies several dominant failure modes relevant to this study [5]. Air leaks are the most common issue, leading to continuous pressure losses and increased compressor load. Persistent leakage forces the compressor to operate near a 100% duty cycle, causing overheating, lubricant degradation, valve fouling, and eventual mechanical failure [2]. Another frequent failure mechanism is desiccant saturation, which allows moisture to enter the braking system, resulting in corrosion of solenoid valves and actuator malfunctions.

These physical degradation processes – elevated temperature, altered pressure dynamics, and extended compressor activity – are reflected in telemetry data. Predictive diagnostics aims to detect latent correlations among these parameters at early degradation stages, while the system remains nominally operational.

Data Analysis Challenges in Diagnostic Tasks. Telemetry data collected from Scania truck onboard systems present several challenges for classical machine learning methods, including high feature dimensionality, multicollinearity, and severe class imbalance.

From a machine learning perspective, the dataset constitutes a high-dimensional classification problem, giving rise to the curse of dimensionality. As the number of features increases, the data space expands exponentially, causing data sparsity. In the 171-dimensional feature space, distances between data points become nearly indistinguishable, reducing the effectiveness of distance-based algorithms such as k-nearest neighbors and requiring substantially more training data than the available ~60,000 observations.

Another critical issue is multicollinearity, evidenced by correlation analyses reported in [5] and [12]. In the APS context, this arises from the physical coupling of system variables: increases in compressor temperature are directly associated with pressure changes and operating time. Additionally, histogram-based features (e.g., ag_000–ag_009) are inherently interdependent, as changes in one bin influence neighboring bins [3].

While multicollinearity does not necessarily reduce overall predictive performance, it increases model sensitivity to noise and complicates feature importance interpretation, as the contribution of correlated variables becomes ambiguous.

Dataset Selection and Justification for Predictive Maintenance Model. One of the main challenges in developing Predictive Maintenance (PdM) systems is the lack of real industrial data. Access to telemetry from complex equipment is limited due to OEM policies, and most publicly available datasets are either synthetic or laboratory-based, which do not reflect the stochastic nature of real-world operations.

A unique resource is the APS Failure and Operational Data for Scania Trucks dataset [11], containing real data from the Air Pressure System (APS) of trucks. APS failure during a trip halts the vehicle, causing significant financial losses. The modeling goal is to predict APS component failures, enabling a shift from reactive to proactive maintenance.

The data are divided into two classes:

- Positive: APS failure;
- Negative: other failures.

The dataset is highly imbalanced: 59,000 negative versus 1,000 positive samples. Attributes are anonymized (e.g., ag_002, cn_005) and represented as histograms by the onboard computer (bins), requiring specific preprocessing to restore statistical properties.

Economic impact of prediction errors:

- False Negative (FN) = 500 units (risk of accident, towing, delivery disruption);
- False Positive (FP) = 10 units (mechanic time to check a healthy vehicle).
- The 50:1 ratio highlights the need to carefully select the most relevant evaluation metrics.

Statistical Analysis and Feature Characteristics. Model performance strongly depends on data quality. Analysis of the Failure target variable confirms extreme class imbalance: 98.3% of records belong to the negative class, and only 1.7% (1,000 records) to the positive class.

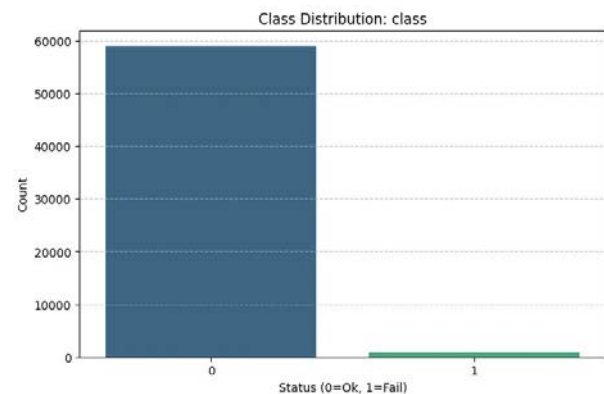


Fig. 4. Histogram of class distribution (0 – normal, 1 – APS failure)

Source: created by the authors.

This imbalance risks the model predicting only class 0, ignoring critical failures.

Data quality assessment revealed a substantial number of missing values (NaN): 171 columns contain at least one missing value, and some features are missing in over 80% of cases, making them unsuitable for modeling.

The high sparsity has a systematic nature due to:

1. Fleet heterogeneity (optional equipment absent in some truck variants);
2. Histogram-based aggregation – missing values often indicate no events in a given bin rather than sensor failure.

Thus, NaNs carry semantic information and require specific imputation rather than simple deletion.

Data Preprocessing Methodology. Due to the characteristics of the Scania APS dataset (high sparsity, anonymized counters, histogram-based features), standard preprocessing methods like row deletion or mean imputation are ineffective. Instead, a heuristic strategy was applied, interpreting missing values (NaN) as informative signals in industrial IoT systems.

Four hypotheses justify this approach [3]:

1. Industrial protocols often transmit data only upon state change or threshold breach; missing records may indicate normal operation.

2. Fleet heterogeneity: NaNs can indicate physically absent sensors in some trucks; dropping these columns would remove crucial differentiation.

3. NaN may signal sensor or wiring failure, representing a predictive feature.

4. Histogram bins: NaN often indicates zero events in the bin.

Implementation: Constant Imputation with marker -1 preserves NaNs as a separate mathematical category, allowing tree-based algorithms to treat them as informative. Normalization was skipped to maintain this marker's semantics, which could be distorted by standard scaling.

Machine Learning Algorithm Selection. The tabular, heterogeneous, and imbalanced nature of the dataset favors tree-based ensemble methods over deep learning or linear models [4, 5, 12]. Ensembles reduce bias and variance:

- Bagging (Random Forest): reduces variance on noisy telemetry.

- Boosting (XGBoost, CatBoost): reduces bias and focuses on rare-class examples, improving sensitivity.

Based on this, three algorithms that demonstrate the best performance on tabular data were selected for further modeling and comparative analysis: Random Forest, XGBoost, and CatBoost.

The Random Forest algorithm, proposed by L. Breiman [13], was chosen as the base model due to its robustness to overfitting and its ability to efficiently handle high-dimensional data without extensive hyperparameter tuning.

Mathematical model: Let $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be the training dataset. The algorithm generates B new bootstrap samples D_b (sampling with replacement). On each sample D_b , a decision tree $f_b(x)$ is trained.

The final classification prediction is obtained by averaging probabilities (for soft voting) or by majority voting:

$$\hat{y} = \arg \max_c \frac{1}{B} \sum_{b=1}^B I(f_b(x) = c), \quad (1)$$

where $I(\cdot)$ is the indicator function, and B is the number of trees in the ensemble.

The criterion for splitting tree nodes is the Gini Impurity. For a node t with class probabilities p_k (where $k \in \{0, 1\}$), the Gini index is calculated as:

$$\text{Gini}(t) = 1 - \sum_{k=0}^1 p_{t,k}^2. \quad (2)$$

Minimizing this index allows the tree to find feature thresholds (e.g., $\text{ag_002} > 0$) that best separate healthy trucks from trucks with defects.

The XGBoost algorithm (eXtreme Gradient Boosting), developed by T. Chen and C. Guestrin [14], is a de facto industry standard for working with tabular data, especially in tasks where accurate classification of rare events is critical.

Mathematical model: The prediction at step t is defined as the sum of the previous prediction and the output of a new tree:

$$\hat{y}_t^{(t)} = \hat{y}_t^{(t-1)} + \eta f_t(x_i), \quad (3)$$

where η is the learning rate.

A key advantage of XGBoost is its specific objective function, which consists of a loss function L and a regularization term Ω .

$$\theta(\phi) = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k). \quad (4)$$

For binary classification tasks (LogLoss), the loss function is:

$$L(y, \hat{p}) = -[y \log(\hat{p}) + (1 - y) \log(a - \hat{p})] \quad (5)$$

The regularization component $\Omega(f)$ is critical for our small set of positive examples, as it penalizes model complexity and prevents overfitting:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (6)$$

where T is the number of leaves in the tree, ω are the leaf weights, and γ and λ are regularization parameters. The

Table 1. Features with the highest share of missing values

Attribute	NaN Count	Missing, %	Attribute	NaN Count	Missing, %
br_000	49 264	82.1%	bl_000	27 277	45.4%
bq_000	48 722	81.2%	bk_000	23 034	38.3%
bp_000	47 740	79.5%	cf_000	14 861	24.7%
bo_000	46 333	77.2%	co_000	14 861	24.7%
cr_000	46 329	77.2%	ch_000	14 861	24.7%
ab_000	46 329	77.2%	ad_000	14 861	24.7%
bn_000	44 009	73.3%	cg_000	14 861	24.7%
bm_000	39 549	65.9%			

Source: created by the authors.

presence of Ω allows XGBoost to generalize better on the test set compared to standard gradient boosting.

Unlike Random Forest, which builds trees in parallel, XGBoost uses Boosting – a sequential construction of the ensemble. Each subsequent tree attempts to correct the errors (reduce the residuals) made by the previous trees.

CatBoost is a high-performance gradient boosting library developed by L. Prokhorenkova [15]. Although our dataset lacks explicit categorical variables (text fields), CatBoost's specific architecture makes it a strong candidate due to its unique ability to handle small samples without overfitting.

Mathematical Model: Unlike standard algorithms that use average gradient values, CatBoost employs Ordered Boosting. To estimate the gradient for the i -th instance, the model only utilizes models trained on instances that precede the i -th instance in a random permutation σ .

In addition, CatBoost builds Oblivious Trees (Symmetric Trees), where the same predicate (splitting condition) is used for all nodes at a given depth level. This can be expressed as a leaf index:

$$Index(x) = \sum_{d=0}^{D-1} 2^d * I(x_{f_d} > c_d), \quad (7)$$

where D is the tree depth, and f_d and c_d are the feature and threshold for depth d , respectively. This structure allows for extremely rapid computation of the prediction $\hat{y}$ during model deployment, which is vital for real-time systems.

Comparative Analysis of Selected Models. The systematic comparison of the chosen algorithms is based on the research by Bentéjac et al. [16] and foundational theoretical works [13-15]. Although all three belong to the tree-based family, they exhibit distinct behaviors:

- **Robustness:** Random Forest (RF) is the least sensitive to outliers due to bagging. In contrast, boosting algorithms may "over-adjust" to noise without proper regularization.

- **Overfitting:** XGBoost is prone to overfitting on small datasets, a challenge CatBoost mitigates via Ordered Boosting, which ensures unbiased gradient estimation [15].

- **Interpretability:** All models support SHAP values, enabling a clear explanation of each sensor's impact on the final diagnosis.

The architectural differences in the context of APS diagnostics are summarized in Table 2.

Table 2. Comparative characteristics of selected classification models

Feature	Random Forest	XGBoost	CatBoost
Ensemble Type	Bagging	Boosting	Boosting
Outlier Sensitivity	Low	Medium	Low
Imbalance Handling	Class Weight	Scale_pos_weight	Class_weights
Overfitting Risk	Low	High (needs tuning)	Low (Ordered Boosting)
Training Speed	Medium	High (GPU)	High
Interpretability	Medium	High (SHAP)	High

Source: created by the authors.

Utilizing these three models provides a balanced approach to APS failure prediction, combining structural stability with high predictive accuracy.

Total Cost: A Domain-Specific Metric. Given the industrial nature of this task, the primary quality criterion is the minimization of operational financial losses. According to the cost matrix defined by Scania CV AB for the Industrial Challenge 2016 [11], the penalties for misclassification are non-equivalent:

- **Cost 1 (False Positive, FP):** Estimated at 10 units. This represents the labor cost for a mechanic to inspect a vehicle that is actually functional (Type I Error).

- **Cost 2 (False Negative, FN):** Estimated at 500 units. This reflects the cumulative losses from an emergency breakdown, including towing, field repairs, and logistics-related reputational damage (Type II Error).

The 1:50 cost ratio acts as a rigid constraint that dictates the optimization strategy. The Total Cost metric is defined by the following formula:

$$\text{Total Cost} = (FP * 10) + (FN * 500). \quad (9)$$

This metric serves as an integral indicator of business efficiency. The optimal model is the one that achieves the minimum Total Cost, regardless of whether it provides the highest global Accuracy or F1-score. This approach forces the algorithm to shift its decision boundary toward the minimization of False Negatives, ensuring maximum reliability for the diagnostic system.

Model Configuration and Training Strategy. To address the extreme class imbalance and structural sparsity of the Scania APS dataset, all models were integrated into a unified processing pipeline. Missing values were imputed with a constant (-1), enabling tree-based architectures to utilize data absence as a predictive feature without introducing the bias inherent in complex interpolation.

The core of the configuration involved adapting loss functions to the minority class:

- **Random Forest:** Implemented an automated weighting heuristic with 100 estimators to stabilize variance.

- **Gradient Boosting (XGBoost & CatBoost):** Employed gradient scaling, where the positive class weight was defined by the negative-to-positive sample ratio. To ensure comparability, both models were synchronized with a maximum tree depth of 6, a learning rate of 0.1, and logarithmic loss optimization.

To mitigate the high risk of overfitting – driven by the limited "Failure" samples (approx. 1,000 records) – a Hold-out Validation strategy was adopted. All scaling and training occurred strictly on the training subset to prevent data leakage.

Regularization and Robustness: Overfitting was further controlled through three primary mechanisms: complexity constraints (limiting tree depth to 6), categorical handling of missing data, and threshold tuning. Rather than maximizing global Accuracy, which favors

the majority class, the decision threshold was optimized to minimize the specific cost function. This ensures the model remains robust and aligned with the economic objectives of the diagnostic system.

Performance Evaluation and Cost Analysis. The primary performance benchmark for this study is the minimization of Type II errors (False Negatives, FN). This prioritization is dictated by the asymmetric cost structure of the Scania APS system, where a missed failure costs 500 units, while a false alarm (False Positive, FP) incurs only 10 units. To evaluate the models, confusion matrices were generated following decision threshold optimization (Fig. 5).

The experimental data reveals distinct predictive behaviors across the three architectures:

- CatBoost: Demonstrated the highest economic efficiency with a total cost of 8,640 units. The model achieved superior Recall, missing only 4 failures across the entire test set. Despite a higher volume of false positives (664) compared to XGBoost, this "cautious" strategy is economically justified, as the cost of additional inspections is significantly lower than the expense of catastrophic system failures.

- XGBoost: Exhibited a tendency toward higher Specificity, producing the fewest false positives (524).

However, in a high-stakes diagnostic environment, this led to an unacceptable increase in missed failures (9 FN). Given the high penalty per miss, total losses rose to 9,740 units, making this model less suitable for safety-critical implementations.

- Random Forest: Yielded the least favorable economic outcome (10,920 units). The model displayed high variance and "noise," generating 792 false alarms. While it missed fewer failures (6 FN) than XGBoost, the excessive burden on service centers significantly reduces its operational profitability.

The results, summarized in Fig. 6, confirm that the CatBoost architecture provides the optimal balance between sensitivity and operational expenditure. By reducing critical misses by more than 50% compared to its closest competitor (4 vs. 9 for XGBoost), CatBoost proves to be the most robust solution for real-time predictive maintenance.

Development of the Diagnostic System Software Prototype. The developed diagnostic system is based on a modular client-server architecture that ensures a clear separation between the Research Environment and the Production Environment. This architectural approach guarantees system scalability and allows for updating decision-making logic and model parameters without modifying the core application's source code.



Fig. 5. Confusion matrices and cost calculation for Random Forest, XGBoost, and CatBoost models
 Source: created by the authors.

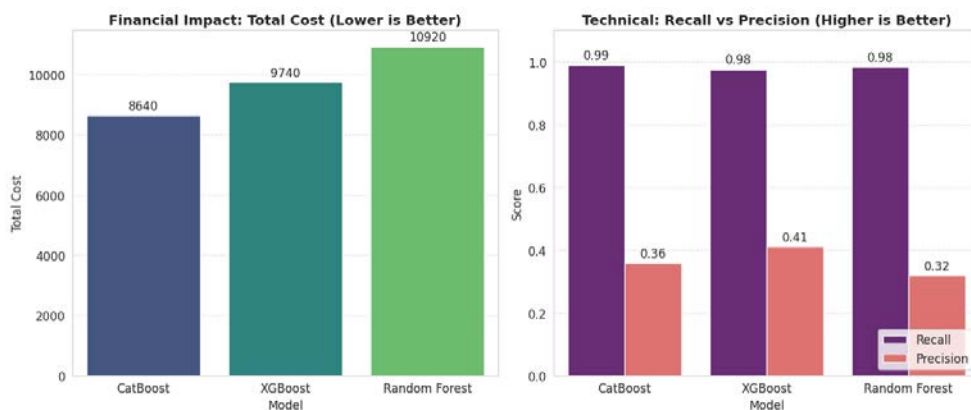


Fig. 6. Comparison of economic efficiency (Total Cost) and technical metrics (Recall/Precision) of the developed models

Source: created by the authors.

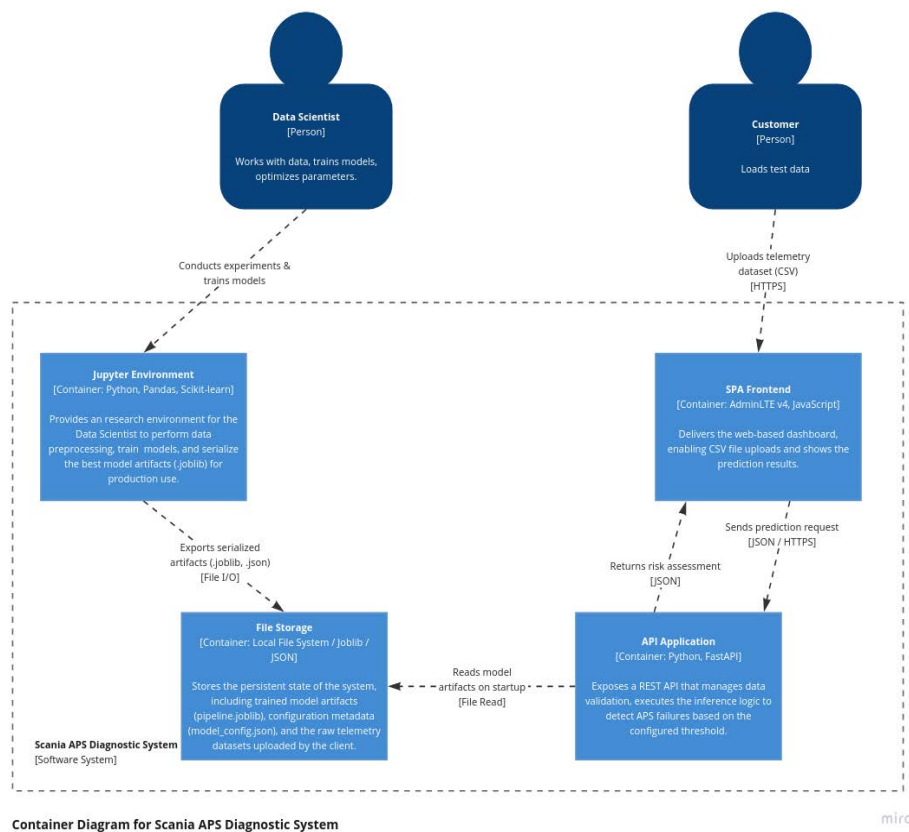


Fig. 7. Software architecture and data flows

Source: created by the authors.

Functionally, the software complex consists of three interconnected modules: an autonomous training module, an inference server, and an operator interface. The information interaction scheme between these components is shown in Fig. 7.

The Offline Training Module (Jupyter Notebook) handles data preprocessing, model training, and validation. It generates portable artifacts: a serialized pipeline (joblib) for consistent data processing and a metadata configuration file (.json) for dynamic threshold adjustments without recompilation.

The Backend Inference Server (FastAPI) provides high-performance, low-latency processing via a RESTful endpoint. It receives sensor data in JSON format, calculates failure probabilities using the pre-trained pipeline, and applies the optimal threshold to issue diagnostic verdicts ("Normal" or "High Risk").

The Frontend Dashboard (AdminLTE v4) provides a web interface for telemetry uploads and interpretation. It features a visual alert system that highlights high-risk vehicles, enabling immediate maintenance prioritization.

Following the MVP implementation, the system was verified using the Scania Trucks test dataset. Given the severe 1:59 class imbalance, an automated evaluation module was integrated to assess performance through both technical and economic metrics.

DISCUSSION OF RESULTS

The study confirms that predictive maintenance for complex mechatronic systems like the Scania APS requires domain-specific financial metrics rather than standard accuracy. CatBoost emerged as the most efficient model, achieving the lowest total cost (8,640 units). Its architectural features – symmetric trees and Ordered Boosting – ensured superior generalization of the rare "Failure" class, missing only 4 cases compared to 9 for XGBoost (a 12.7% difference in financial efficiency).

Unlike previous studies [2, 4] that focused on global accuracy, this research emphasizes cost-sensitive learning. Adapting the loss function to a 1:50 cost ratio and implementing Threshold Tuning proved vital for aligning the model with the business realities of logistics.

Validity of Hypotheses:

- Minimizing Total Cost is a more objective success criterion than Accuracy for imbalanced industrial data.
- Treating missing values (NaN) as an informative category improved forecast stability.
- Gradient boosting, particularly CatBoost, outperformed Random Forest in identifying minority class failures.

Thus, the developed model and the FastAPI-based software prototype serve as a viable technological foundation for implementing a predictive maintenance strategy in real-world Scania truck operating conditions.

CONCLUSIONS

In the course of this study, an important scientific and applied problem was addressed: the development of a decision-support model for heavy-duty vehicle diagnostics that minimizes the financial risks of logistics companies.

The main results of the research are as follows:

1. Justification of the Total Cost Metric: It has been proven that for critical diagnostic systems, standard metrics such as Accuracy and F1-score are insufficient. The implementation of a domain-specific cost function with a 1:50 ratio allowed the model training process to adapt to the real-world economic conditions of Scania truck operations.

2. Selection of the Optimal Architecture: Based on comparative analysis, the CatBoost algorithm demonstrated the highest efficiency. Leveraging symmetric tree structures and the Ordered Boosting mechanism, the model achieved the highest Recall, reducing the number of missed failures by more than half compared to

XGBoost (4 vs. 9 cases). This resulted in a minimum total cost of 8,640 units.

3. Development of a Regularization Strategy: A comprehensive approach to handling data sparsity and class imbalance was implemented, including the interpretation of missing values as distinct features and probability threshold optimization. This ensured model robustness against overfitting on small samples of the "Failure" class.

4. Software Implementation: A modular diagnostic system prototype was developed using the FastAPI framework. The architectural decoupling of the training and inference processes allows for rapid deployment and dynamic model updates without service interruption.

The results of this study are of practical significance for the development of intelligent transport monitoring systems. They facilitate the transition from costly scheduled maintenance strategies to an efficient predictive service model, reducing losses from emergency breakdowns by an estimated 10–15%.

REFERENCES

- [1] Prytz, R., Nowaczyk, S., Rögnvaldsson, T., & Byttner, S. (2015). Predicting the need for vehicle compressor repairs using maintenance logs and logged vehicle data. In *Proceedings of the 18th IEEE International Conference on Intelligent Transportation Systems (ITSC 2015)*. P. 872–877. IEEE. <https://doi.org/10.1109/ITSC.2015.143>
- [2] Costa, B. S. J., & Nascimento, N. M. (2016). IDA 2016 industrial challenge: Using machine learning for predicting failures in air pressure systems of Scania trucks. In *Advances in Intelligent Data Analysis XV*. P. 381–386. Springer. https://doi.org/10.1007/978-3-319-46349-0_36
- [3] Cerquitelli, T., Quercia, D., Apiletti, D., Baralis, E., & Venturini, L. (2019). Predictive maintenance of air conditioning systems using supervised machine learning. In *Proceedings of the 2019 IEEE International Conference on Big Data (Big Data)*. P. 5906–5915. IEEE. <https://doi.org/10.1109/BigData47090.2019.9006552>
- [4] Ghorbani, R., & Ghousi, R. (2020). Predicting and categorizing air pressure system failures in Scania trucks using machine learning. In *Proceedings of the International Conference on Industrial Engineering and Operations Management*. P. 2807–2816.
- [5] Weiss, G. M., McCarthy, K., & Zabar, B. (2018). Cost-sensitive learning for predictive maintenance. In *Proceedings of the Daily Dataset Challenge*.
- [6] Zonta, T., da Costa, C. A., da Rosa Righi, R., de Lima, M. J., da Trindade, E. S., & Li, G. P. (2020). Predictive maintenance in the Industry 4.0: A systematic literature review. *Computers & Industrial Engineering*. Vol. 150. P. 106889. <https://doi.org/10.1016/j.cie.2020.106889>
- [7] IVC Technologies. (2017, August 29). Reactive, preventive, predictive maintenance. <https://ivctechnologies.com/2017/08/29/reactive-preventive-predictive-maintenance/>
- [8] Kumar, A. (2024, July 15). Batch processing vs stream processing. K21Academy. <https://k21academy.com/azure-data/batch-processing-vs-stream-processing/>
- [9] Khattach, O., Moussaoui, O., & Hassine, M. (2025). End-to-end architecture for real-time IoT analytics and predictive maintenance using stream processing and ML pipelines. *Sensors*. Vol. 25(9). P. 2945. <https://doi.org/10.3390/s25092945>
- [10] Scania CV AB. (2016). APS failure at Scania trucks data set. UCI Machine Learning Repository. <https://archive.ics.uci.edu/dataset/421/aps+failure+at+scania+trucks>
- [11] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*. Vol. 35. P. 507–520. <https://doi.org/10.48550/arXiv.2207.08815>
- [12] Breiman, L. (2001). Random forests. *Machine Learning*. Vol. 45(1). P. 5–32. <https://doi.org/10.1023/A:1010933404324>
- [13] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. P. 785–794. <https://doi.org/10.1145/2939672.2939785>
- [14] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*. Vol. 31. P. 6638–6648. <https://doi.org/10.48550/arXiv.1706.09516>
- [15] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*. Vol. 54(3). P. 1937–1967. <https://doi.org/10.1007/s10462-020-09896-5>
- [16] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*. Vol. 21(9). P. 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- [1] Prytz, R., Nowaczyk, S., Rögnvaldsson, T., & Byttner, S. (2015). Predicting the need for vehicle compressor repairs using maintenance logs and logged vehicle data. In *Proceedings of the 18th IEEE International Conference on Intelligent Transportation Systems (ITSC 2015)*. P. 872–877. IEEE. <https://doi.org/10.1109/ITSC.2015.143>
- [2] Costa, B. S. J., & Nascimento, N. M. (2016). IDA 2016 industrial challenge: Using machine learning for predicting failures in air pressure systems of Scania trucks. In *Advances in Intelligent Data Analysis XV*. P. 381–386. Springer. https://doi.org/10.1007/978-3-319-46349-0_36
- [3] Cerquitelli, T., Quercia, D., Apiletti, D., Baralis, E., & Venturini, L. (2019). Predictive maintenance of air conditioning systems using supervised machine learning. In *Proceedings of the 2019 IEEE International Conference on Big Data (Big Data)*. P. 5906–5915. IEEE. <https://doi.org/10.1109/BigData47090.2019.9006552>
- [4] Ghorbani, R., & Ghousi, R. (2020). Predicting and categorizing air pressure system failures in Scania trucks using machine learning. In *Proceedings of the International Conference on Industrial Engineering and Operations Management*. P. 2807–2816.
- [5] Weiss, G. M., McCarthy, K., & Zabar, B. (2018). Cost-sensitive learning for predictive maintenance. In *Proceedings of the Daily Dataset Challenge*.
- [6] Zonta, T., da Costa, C. A., da Rosa Righi, R., de Lima, M. J., da Trindade, E. S., & Li, G. P. (2020). Predictive maintenance in the Industry 4.0: A systematic literature review. *Computers & Industrial Engineering*. Vol. 150. P. 106889. <https://doi.org/10.1016/j.cie.2020.106889>
- [7] IVC Technologies. (2017, August 29). Reactive, preventive, predictive maintenance. <https://ivctechnologies.com/2017/08/29/reactive-preventive-predictive-maintenance/>
- [8] Kumar, A. (2024, July 15). Batch processing vs stream processing. K21 Academy. <https://k21academy.com/azure-data/batch-processing-vs-stream-processing/>
- [9] Khattach, O., Moussaoui, O., & Hassine, M. (2025). End-to-end architecture for real-time IoT analytics and predictive maintenance using stream processing and ML pipelines. *Sensors*. Vol. 25(9). P. 2945. <https://doi.org/10.3390/s25092945>
- [10] Scania CV AB. (2016). APS failure at Scania trucks data set. UCI Machine Learning Repository. <https://archive.ics.uci.edu/dataset/421/aps+failure+at+scania+trucks>
- [11] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*. Vol. 35. P. 507–520. <https://doi.org/10.48550/arXiv.2207.08815>
- [12] Breiman, L. (2001). Random forests. *Machine Learning*. Vol. 45(1). P. 5–32. <https://doi.org/10.1023/A:1010933404324>
- [13] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. P. 785–794. <https://doi.org/10.1145/2939672.2939785>
- [14] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*. Vol. 31. P. 6638–6648. <https://doi.org/10.48550/arXiv.1706.09516>
- [15] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*. Vol. 54(3). P. 1937–1967. <https://doi.org/10.1007/s10462-020-09896-5>
- [16] He, H., & García, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*. Vol. 21(9). P. 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>

© Бойко Н. І., Боднар Ю. В.

Дата першого надходження статті до видання: 04.02.2026

Дата прийняття статті до друку після рецензування: 02.03.2026

Дата публікації (оприлюднення) статті: 22.05.2026



Стаття поширюється на умовах ліцензії відкритого доступу (CC BY 4.0)