Cost-Sensitive Deep Learning for Scania Component X: Minimising Operational Cost via Asymmetric Threshold Optimisation

Abdelhakim Mraihi[1], Valeriu Dimidov[2], Raoof Doorshi[1], Reza Khoshkangini[1]

[1] *Malmö University, Malmö, Sweden*
hakimmraihi1@gmail.com, r.doorshi@gmail.com, reza.khoshkangini@mau.se
[2] *SnT, Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Esch-sur-Alzette, Luxembourg*
valeriu.dimidov@uni.lu

ABSTRACT

Maintenance decisions in industrial fleets must balance the cost of unnecessary interventions against the higher cost of missed failures. This paper presents a cost-sensitive deep learning approach for the Scania Component X benchmark. Rather than predicting the original five degradation classes directly, the problem is reformulated as a binary task that identifies whether a vehicle is healthy or at risk based on recent operational data. The predicted risk score is then converted into a maintenance decision through asymmetric threshold optimisation under the official Scania 5×5 cost matrix. Three temporal deep learning models—CNN, Transformer, and Temporal Convolutional Networks—are evaluated under the same cost-aware training and decision setting. Results on real data from thousands of heavy-duty trucks show how cost-sensitive learning affects the trade-off between failure avoidance, maintenance workload, and total operational cost.

1. INTRODUCTION

Heavy-duty vehicles are complex industrial systems operating under a wide range of configurations, environments, and usage conditions. In such systems, component downtime may originate from multiple sub-components that degrade or fail for different reasons over time. The failure to anticipate component breakdowns can lead to increased maintenance costs, unplanned vehicle downtime, safety risks for customers, and reduced service reliability, ultimately affecting customer satisfaction and brand value. Consequently, the accurate and timely prediction of component failures has become a central objective in modern predictive maintenance strategies for vehicle manufacturers.

Driven by this objective, much research has investigated fault prediction, reliability analysis, and degradation modelling in the automotive and industrial domains. Early approaches relied on statistical and stochastic degradation models to estimate failure risk and remaining useful life from historical usage and failure data Ding and Fang (2017); Man and Zhou (2018). With the increasing availability of sensors and operational data, Machine Learning (ML) techniques have been widely adopted to capture more complex and non-linear degradation patterns. Comprehensive reviews confirm that classical ML and Deep Learning (DL) methods significantly improve predictive performance across a wide range of machinery fault diagnosis and prognostic applications Deldari, Loke, and Zaslavsky (2023); Lei et al. (2020).

In the automotive domain, machine learning has also been applied to vehicle behaviour modelling and warranty claim prediction using large-scale operational vehicle data Khoshkangini, Mashhadi, Tegnered, and Rognvaldsson (2022); Khoshkangini, Pashami, and Nowaczyk (2019).

More recently, DL models have shown strong potential for Predictive Maintenance (PdM) by directly learning temporal and multivariate representations from operational data. Contrary to statistical methods that rely on predefined distributions (e.g. Weibull) that may not reflect reality, DL captures complex non-linear patterns effectively Deldari et al. (2023). Convolutional, recurrent, and attention-based architectures have been applied to component degradation modelling, Remaining Useful Life (RUL) estimation, and failure prediction in industrial systems Mashhadi et al. (2019). At the same time, several studies highlight that predictive accuracy alone is insufficient for real-world deployment, as maintenance decisions are governed by strongly asymmetric business costs Shah, Wilder, Perrault, and Tambe (2024). In the industrial fleet operations, the cost of missing an imminent failure typically exceeds the cost of unnecessary preventive maintenance by a large margin, motivating the use of cost-sensitive learning and decision-oriented evaluation frameworks Chehade et al. (2022); Rengasamy, Jafari, and Rothwell (2020); Zhang, Wang, and Li (2023).

The release of the *Scania Component X dataset* Kharazian, Lindgren, Magnússon, Steinert, and Andersson Reyna (2025) provides a unique opportunity to study PdM under realistic industrial conditions. The dataset contains real-world multivariate time-series data collected from thousands of heavy-duty trucks and includes an asymmetric cost matrix explicitly designed to reflect industrial maintenance priorities . Recent studies have benchmarked classical machine learning, contextual models, and graph-based approaches on this dataset, demonstrating that models achieving high accuracy do not necessarily yield the lowest operational cost Carpentier, De Temmerman, and Verbeke (2024); Dintén, Zorrilla, Veloso, and Gama (2025); Parton, Fois, Vegliò, Metta, and Gregnanin (2024); Yang and Iqbal (2025). These findings reinforce the importance of cost-aware and decision-driven evaluation protocols.

Despite these advances, there is still a limited understanding of how modern temporal DL architectures behave when trained and evaluated under a strictly business-driven framework. In particular, few studies provide a systematic comparison of Deep Neural Network (DNN)s while jointly analysing cost efficiency, failure avoidance, and maintenance workload under identical operational constraints. Recent work on risk-aware decision-making further emphasises the need to explicitly optimize maintenance decisions rather than relying on fixed classification thresholds Johnson and Khoshgoftaar (2021); Xu and Zhang (2026).

In this work, we address these gaps by developing a cost-sensitive maintenance decision approach aligned with Scania's operational cost structure. We reformulate multi-stage degradation into a binary operational decision, while preserving full compatibility with the official asymmetric evaluation protocol. Multiple temporal DL architectures are evaluated under identical conditions, and a statistical analysis is conducted to compare their economic cost, failure avoidance, and maintenance burden. By combining cost-sensitive modelling with systematic statistical evaluation on real fleet data, this study aims to support more reliable and economically meaningful maintenance decisions in industrial vehicle operations.

This study is guided by the following research questions:

- **RQ1:** How do different temporal deep learning architectures behave under a cost-aware maintenance decision setting?
- **RQ2:** How does cost-sensitive optimisation influence failure avoidance and the resulting maintenance workload in real fleet operations?
- **RQ3:** Which modelling approach provides the most favourable trade-off between operational cost and failure risk on real-world fleet data?

The remainder of this manuscript is organized as follows. Section 2 presents the methodology and experimental protocol. Section 3 reports the experimental results and compares the evaluated models under the Scania cost matrix. Section 4 discusses the main findings and their implications. Finally, Section 5 concludes the manuscript and summarizes the main contributions.

2. METHODOLOGY

2.1. Dataset

In our experiments, we used the Scania Component X dataset provided by Kharazian et al. (2025) for the IDA 2024 Challenge. The dataset contains real-world multivariate time series collected from a fleet of 33,641 trucks. It is divided into training (70%), validation (15%), and test (15%) partition as detailed in Table 1, which provides the number of vehicles and readouts for each segment.

Partition	Vehicles	Readouts
Train	23550	1122452
Validation	5046	196227
Test	5045	198140

Table 1. Number of Vehicles and Time Series Observations in the train, validation, and test partitions

The data includes operational readouts related to an anonymized engine component (on-board data) as well as vehicle specifications and repair records (off-board data).

2.1.1. Vehicle Specifications

The vehicle specification data provide static information about each truck in the fleet. These data are represented by eight anonymized categorical attributes, denoted as Spec_0 to Spec_7. Although the original meaning of the attributes is not disclosed, they encode relevant configuration characteristics of the vehicles, such as possible differences in engine configuration, drivetrain layout, or other structural properties. Each variable assumes a finite set of textual categories, represented in the dataset by values such as "Cat0", "Cat1", and so forth.

Feature	Cardinality
Spec_0	3
Spec_1	29
Spec_2	21
Spec_3	4
Spec_4	2
Spec_5	5
Spec_6	17
Spec_7	9

Table 2. Number of unique values for specification feature

Table 2 reports the number of distinct categories observed for each specification feature. These variables complement the operational readouts by adding vehicle-level contextual information. From a predictive maintenance perspective, such

static descriptors may be useful because trucks with different configurations can experience different usage patterns, degradation dynamics, and failure probabilities.

2.1.2. Operational Readouts

The operational readout data consist of multivariate temporal measurements collected from the monitored vehicles. Overall, the readouts include 107 anonymized features associated with the operation of Component X. These features can be grouped into two main families: single-counter variables and histogram-based variables.

The single-counter variables are cumulative measurements that evolve over time. Since they represent accumulated quantities, their values are generally expected to follow non-decreasing trajectories.

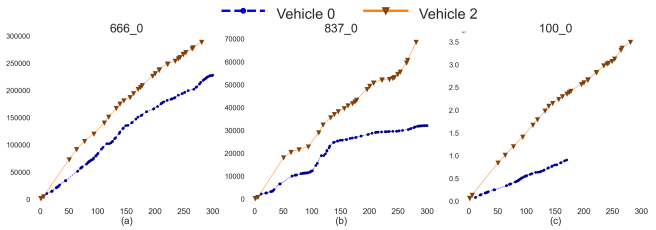


Figure 1. Comparison of Features 666_0, 837_0, 100_0 between Vehicles 0 and 2

Figure 1 illustrates the behavior of selected counter features, namely 666_0, 837_0, and 100_0, for vehicles 0 and 2 over a horizon of 300 time steps. The comparison highlights two important properties of the data. First, the counters are defined on different numerical scales. Second, the vehicles are not observed at regular intervals, resulting in irregular sampling patterns. Moreover, the rate of increase can vary substantially across vehicles, suggesting differences in usage intensity or operating conditions. For instance, feature 666_0 shows an approximately linear trend for both vehicles, but with different slopes. By contrast, feature 837_0 exhibits a less regular evolution, especially in the initial part of the series. In subplot (c), the sequence for vehicle 0 terminates earlier, around time step 200, due to missing values that affect the completeness of the corresponding time series.

In addition to the single counters, the dataset contains six groups of histogram-based features. Each group originates from an underlying counter and is split into multiple bins, where each bin summarizes the amount of operation performed under a specific condition. As these quantities are also accumulated over time, histogram-based features typically display non-decreasing behavior as well.

Figure 2 summarizes the main characteristics of the operational readouts across the training, validation, and test partitions. The three partitions show comparable distributions in terms of observation window, number of readouts, and sampling frequency. Nevertheless, the time series remain highly heterogeneous: vehicles differ both in the number of available observations and in the temporal distance between consecutive readouts. This irregularity directly affects the amount of information available for each vehicle and complicates the modeling of the degradation process.

2.1.3. Time-to-Event

The time-to-event data describe the maintenance outcome associated with Component X. For each vehicle in the training partition, the dataset specifies whether a repair of Component X occurred during the observation period through the variable in_study_repair. When such an event is observed, the variable length_of_study_time_step indicates the time step at which the repair took place.

This information is used to reconstruct the remaining useful life of Component X for the training vehicles and, consequently, to assign degradation labels to their operational readouts. However, the time-to-event file is not provided for the validation and test partitions. For these partitions, the dataset only includes the class label associated with the last available operational readout of each vehicle. The procedure used to derive class labels from the time-to-event information is described in the following section.

2.2. Readings Classification and Failure Analysis

Class	RUL
0	> 48
1	$[48, 24)$
2	$[24, 12)$
3	$[12, 6)$
4	$[6, 0)$

Table 3. Class labeling logic

Operational readouts from trucks are categorized into five classes based on RUL of the component, calculated as $RUL = failure\_time - time\_step$. The class labeling follows the scheme presented in Table 3. Readouts sampled when Component X is healthy are labeled with class 0 and are associated with a $RUL > 48$. When the degradation process begins, readouts are labeled with $\{1, 2, 3, 4\}$ according to the RUL time windows of [48, 24), [24, 12), [12, 6), [6, 0), respectively.

The labeling rules assume that once the degradation process starts, it is irreversible and continues until the end of life of the vehicle. As the trucks approach the failure event, the assigned class reflects an increasing risk of failure. However, when the historical data of a truck is right censored because the failure event did not happen during the study, all the readings of the vehicle are assigned to class 0. This follows the dataset setup, where degradation is assumed to continue once it starts.

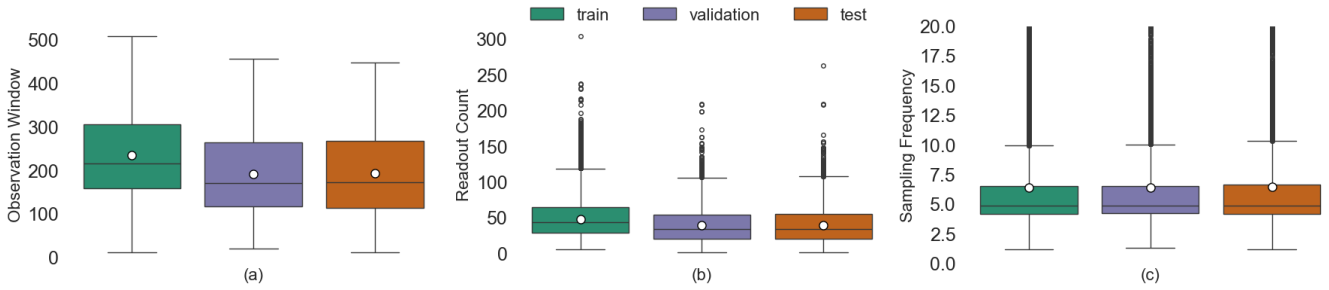


Figure 2. Descriptive statistics of the dataset (training, validation and test)

The process of labeling readouts in relation to the end-of-study and failure events is illustrated in Figure 3.

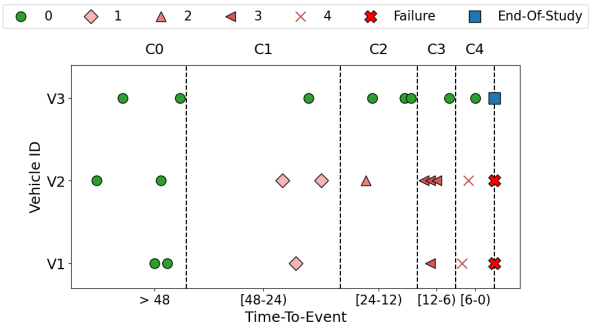


Figure 3. Readout classification by time-to-event

Vehicles V1 and V2 experience failures within the study period, so their readouts that occur within 48 time steps prior to the failure event are labeled according to the rules 1-4 outlined in Table 3. In contrast, vehicle V3 operates without any malfunctions throughout the study duration. Therefore, all observations related to V3 are categorized as 0.

The fleet exhibits an imbalance between vehicles that required repairs during the study period (faulty) and those that did not (healthy).

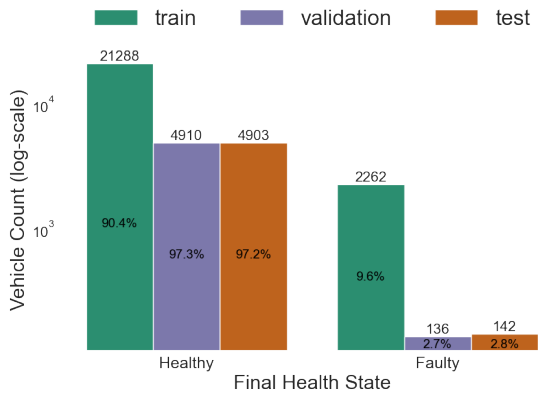


Figure 4. Distribution of Healthy vs. Faulty vehicles across training, validation and test datasets

In Figure 4, we can see that only a small proportion of trucks required repairs to component X. This imbalance is particularly pronounced in the validation and test partitions, where more than 97% of the vehicles are healthy. The distribution of healthy and faulty vehicles is similar in these two partitions, while the training set has a slightly higher proportion of faulty vehicles (9.6%).

Ultimately, it is important to note that the time series are irregular, which means that the faulty measurements do not always fall into the categories that capture the degradation process of component X. For example, in Figure 3 no readout of class 2 is registered for V1 despite the vehicle experiencing a faulty event.

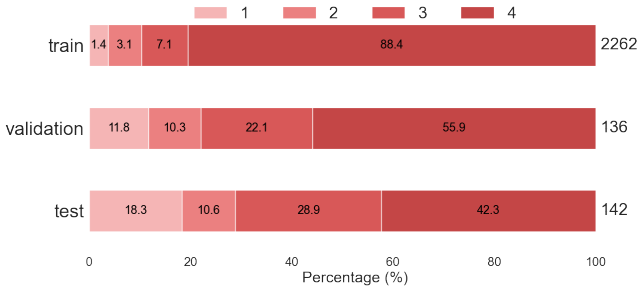


Figure 5. Distribution faulty classes across last readouts

As seen in Figure 5, the last readout of the training dataset presents a strong bias towards the most severe degradation category. In contrast, the validation and test datasets present a more even distribution, potentially offering a more accurate assessment of the model's performance.

2.3. Problem Definition

We consider a fleet of n vehicles, denoted as $\mathcal{V} = \{V_0, V_1, \ldots, V_{n-1}\}$, monitored using on-board sensors during their operational lifetime. For each vehicle V_i, the collected data form an irregular multivariate time series $X_i \in R^{T_i \times M}$, where T_i denotes the number of recorded readouts and M the number of operational features.

Each readout is associated with a degradation label

$y_t \in \{0, 1, 2, 3, 4\}$, defined according to the RUL of the monitored component, following the official Scania Component X labeling scheme. The five classes correspond to progressively shorter time windows before failure, with class 0 indicating healthy operation and class 4 representing imminent failure.

Operational Objective. Although the dataset is annotated using a five-class degradation taxonomy, the practical maintenance objective is binary: deciding whether a vehicle should be flagged for preventive maintenance at the current time. In industrial fleet operations, maintenance actions are discrete and fail-safe, and the economic consequences of missed failures are substantially higher than those of unnecessary inspections.

Binary Risk Reformulation. To align learning with operational decision-making, we reformulate the problem as a binary failure-risk estimation task. All failure-prone degradation states $\{1, 2, 3, 4\}$ are merged into a single *At-Risk* category, while class 0 is retained as *Healthy*. Each model therefore learns a probabilistic mapping

$$f_{\text{risk}} : R^{T \times M} \rightarrow [0, 1]$$

where $\hat{p}_i = f_{\text{risk}}(X_i)$ represents the estimated probability that vehicle V_i is at risk of imminent component failure.

Fail-Safe Decision Mapping. At deployment time, probabilistic outputs are converted into maintenance decisions using a threshold θ. To ensure full compatibility with the official Component X evaluation protocol, binary predictions are mapped back to the five-class label space using a fail-safe rule:

$$\hat{y}_i(\theta) = \begin{cases} 4, & \hat{p}_i \geq \theta, \\ 0, & \hat{p}_i < \theta. \end{cases}$$

This mapping preserves the asymmetric cost structure defined by Scania, where false-negative errors incur substantially higher penalties than false positives.

Evaluation Protocol. Model performance is evaluated exclusively using the official asymmetric 5×5 cost matrix applied to the mapped predictions. This ensures that, although learning is performed in a binary risk space, all reported results remain directly comparable to prior work on the Scania Component X dataset and adhere strictly to the benchmark's fail-safe evaluation policy.

2.4. Workflow

Figure 6 provides a high-level overview of the proposed architecture. Phase 1 prepares temporal vehicle data, Phase 2 performs cost-sensitive failure risk modelling using deep neural networks, and Phase 3 applies decision-theoretic threshold optimisation and statistical analysis to derive business-optimal maintenance decisions.

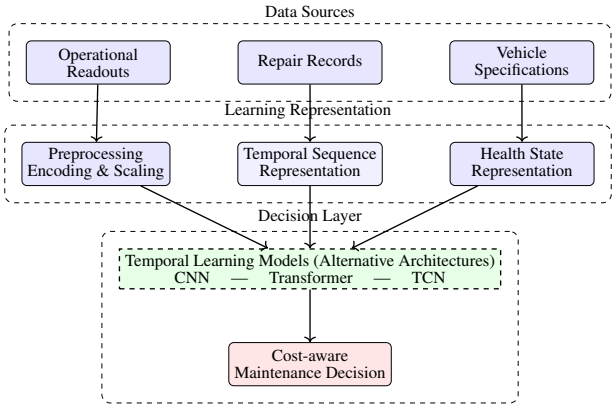


Figure 6. Conceptual architecture of the prposed workflow.

Predictive maintenance on the Scania Component X dataset has previously been approached using contextual ML models, survival analysis, and graph-based techniques Carpentier et al. (2024); Parton et al. (2024); Yang and Iqbal (2025). These studies follow the official fail-safe evaluation protocol introduced with the dataset Kharazian et al. (2025). The protocol defines five degradation classes and employs an asymmetric cost matrix designed to reflect Scania's operational risk structure.

To ensure full comparability with prior work, we adopt the same fail-safe evaluation strategy. At the same time, we introduce a cost-aware learning and threshold optimization framework that embeds business priorities directly into model training and deployment.

2.5. Sequence Construction and Feature Processing

Each vehicle is associated with an irregular sequence of multivariate operational readouts. For every vehicle, we extract the last 30 time steps of available data, with shorter sequences zero-padded at the beginning. Operational readouts are augmented with eight categorical vehicle specifications, which are label-encoded and appended to the temporal representation.

To reduce noise and ensure tractable modelling, we apply:

- **Variance-based feature selection** to retain the 64 most informative operational variables,
- **Median imputation** for missing values,
- **Standard scaling** applied globally across the flattened sequences.

Each training instance therefore corresponds to a fixed-size (30×64) matrix representing the recent operational trajectory

of a vehicle.

2.6. Cost-Aware Failure Risk Modelling

The Component X dataset defines a five-class degradation taxonomy with an associated asymmetric cost structure designed for maintenance decision evaluation Kharazian et al. (2025), as presented in Equation 1. In this setting, missed failures (False Negatives) incur substantially higher penalties than unnecessary preventive maintenance actions (False Positives).

$$\mathrm{COST}_{5\times5} = \begin{bmatrix} 0 & 7 & 8 & 9 & 10 \\ 200 & 0 & 7 & 8 & 9 \\ 300 & 200 & 0 & 7 & 8 \\ 400 & 300 & 200 & 0 & 7 \\ 500 & 400 & 300 & 200 & 0 \end{bmatrix} \tag{1}$$

Nevertheless, we adopt a binary reformulation aligned with industrial maintenance practice rather than training directly on the full five-class cost structure. Specifically, we define a mapping $\mathcal{M}$ such that:

$$\mathcal{M}(y_i) = \begin{cases} 0 & \text{if } y_i = 0 \text{ (Healthy)} \\ 1 & \text{if } y_i \in \{1,2,3,4\} \text{ (At-Risk)} \end{cases} \tag{2}$$

This reformulation simplifies the learning process, stabilises optimisation, and directly supports binary operational decision-making (i.e., whether to inspect the vehicle or not) Dintén et al. (2025).

Cost asymmetry is incorporated during training through a moderated cost-sensitive loss, implemented as weighted binary cross-entropy (WBCE) defined in Equation 3:

$$\mathcal{L}_{WBCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[w_1 \cdot y_i \log(\hat{p}_i) + w_0 \cdot (1-y_i)\log(1-\hat{p}_i)\right] \tag{3}$$

where $y_i \in \{0,1\}$ is the binary ground-truth label, $\hat{p}_i$ is the predicted probability of the vehicle belonging to the At-Risk class, and the class weights are set to $w_1 = 10$ and $w_0 = 1$. This weighting ratio reflects the high penalty of false negatives while avoiding the numerical instability often associated with the dataset's extreme 50:1 cost ratios. Each model eventually outputs a probabilistic estimate $\hat{p}(\text{At-Risk})$, which serves as the basis for the threshold-driven decision process described in the following section.

2.7. Decision-Theoretic Threshold Optimization

During deployment, predicted failure risks are converted into binary maintenance decisions using an operating threshold θ. A vehicle is recommended for preventive maintenance when the predicted probability $\hat{p}(\text{At-Risk}) \geq \theta$. To remain compatible with the official Component X protocol, these binary predictions are mapped back to the five-class space using a fail-safe rule:

$$\hat{y}_i(\theta) = \begin{cases} 4, & \hat{p}_i \geq \theta, \\ 0, & \hat{p}_i < \theta. \end{cases} \tag{4}$$

Given the true labels $y_i \in \{0,\ldots,4\}$, the fleet-wide operational cost is defined as:

$$C(\theta) = \sum_{i=1}^{N} \mathrm{COST}_{5\times5}\left(y_i, \hat{y}_i(\theta)\right). \tag{5}$$

Because the total cost $C(\theta)$ only changes when θ crosses a predicted probability, the optimal threshold satisfies the following objective:

$$\theta^* = \arg\min_{\theta \in \{\hat{p}_1,\ldots,\hat{p}_N\}} C(\theta). \tag{6}$$

We approximate this search over $\theta \in [0.01, 0.50]$ with a step size of 0.01. As established by the extreme cost asymmetry of the dataset, the theoretical Bayes threshold is defined as:

$$\theta_{\text{Bayes}} = \frac{10}{10+500} \approx 0.02. \tag{7}$$

This value indicates a strongly risk-averse operating regime. In practice, however, the final operating threshold θ^* is obtained via post-hoc optimization on the validation set according to Equation 6. To ensure the model remains aligned with business objectives while maintaining training stability, model selection employs early stopping based on the validation cost (Eq. 5) calculated at a fixed reference threshold of $\theta = 0.5$. This two-stage approach ensures that deployment is perfectly aligned with the minimum achievable cost under the full 5×5 matrix.

2.8. Temporal Learning Architectures

The proposed solution is primarily based on a Transformer-based temporal encoder designed to capture long-range dependencies in multivariate operational sequences. To assess its effectiveness, we compare it with two widely used temporal modelling architectures—Convolutional Neural Networks (CNN) and Temporal Convolutional Networks (TCN)—which serve as baseline models under the same training and evaluation protocol. A detailed schematic of the model architectures and their shared cost-aware training and decision logic is provided in Fig. 6.

Convolutional Neural Network: The Convolutional Neural Network (CNN) processes (30×64) operational sequences through three 1D convolutional blocks with channel depths

of 64, 128, and 256. Each block consists of convolution, batch normalisation, ReLU activation, max-pooling, and a dropout. Adaptive global average pooling compresses the temporal dimension into a fixed-size representation, which is passed to a two-layer fully connected classifier producing binary Healthy/At-Risk logics. With approximately 170k trainable parameters, the CNN serves as a lightweight and computationally efficient baseline.

Transformer-Based Temporal Encoder: The Transformer architecture projects each time step into a 128-dimensional latent space and augments the sequence with learned positional encodings to preserve temporal order. A stack of three Transformer encoder layers, each comprising multi-head self-attention, feed-forward sublayers, residual connections, and layer normalization, captures long-range temporal dependencies and cross-sensor interactions. Global average pooling aggregates the contextualized sequence into a single embedding, followed by a two-layer classifier that outputs binary Healthy/At-Risk logits.

Temporal Convolutional Network: The Temporal Convolutional Network (TCN) employs three residual temporal blocks with dilated causal convolutions (dilations 1, 2, and 4), enabling an exponentially expanding receptive field over the 30-step input horizon. Each block combines dilated convolutions, batch normalization, ReLU activation, dropout, and residual connections to stabilize training. Adaptive average pooling collapses the temporal dimension, and a two-layer classifier produces binary outputs. This architecture provides a compromise between long-range temporal modelling capacity and computational efficiency.

Across all architectures, models are trained with the same cost-sensitive objective and feed into the shared threshold optimization and fail-safe evaluation procedure illustrated in Fig. 6.

2.9. Training Configuration and Evaluation Protocol

All models are trained using the same optimisation setup: AdamW with learning rate 1×10^{-4}, batch size 32, gradient clipping with maximum norm 1.0, and weighted cross-entropy loss. Early stopping is applied based on validation-set operational cost.

Evaluation strictly follows the official Component X fail-safe protocol. The primary metric is total fleet-level operational cost. Secondary diagnostics include false-negative and false-positive counts, balanced accuracy, and maintenance rate. This unified protocol ensures strict comparability across architectures and with all published Component X benchmarks.

2.10. Evaluation Protocol and Metrics

All models are evaluated using the official fail-safe protocol defined for the Scania Component X dataset Kharazian et al. (2025). Model outputs are binary risk predictions indicating whether a vehicle snapshot is classified as *at-risk* or *healthy*. These predictions are subsequently mapped to class 4 (at-risk) or class 0 (healthy) before applying the full asymmetric 5×5 cost matrix.

Model thresholds are selected by minimizing total validation-set operational cost under the asymmetric cost matrix. Early stopping is also governed by validation cost rather than training loss to ensure alignment with business objectives.

Operational performance is primarily assessed using total cost and average cost per vehicle. To provide deeper insight into model behaviour, several complementary diagnostic metrics are reported. Failure prevention capability is quantified using the *False Negative Rate (FNR)* as shown in Equation 8:

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}, \tag{8}$$

where FN and TP denote false negatives and true positives, respectively.

Maintenance workload is measured using the *Maintenance Rate*, defined as the proportion of vehicles flagged for preventive maintenance. To relate this workload to the underlying failure prevalence, we compute the *Maintenance Lift Index* defined in Equation 9:

$$\mathrm{Lift} = \frac{\mathrm{Maintenance\ Rate}}{\mathrm{True\ Failure\ Rate}}. \tag{9}$$

Economic efficiency is further analysed using the *Cost per True Positive (Cost/TP)*, defined in Equation 10, which reflects the average operational cost incurred for each correctly identified failure:

$$\mathrm{Cost/TP} = \frac{\mathrm{Total\ Operational\ Cost}}{\mathrm{TP}}. \tag{10}$$

Finally, relative cost reduction is reported with respect to two benchmark strategies—predicting all vehicles as healthy and predicting all as at-risk— as recommended in prior Component X studies Kharazian et al. (2025).

3. RESULTS

Table 4 presents a comprehensive comparison of the evaluated temporal deep learning architectures under the asymmetric 5×5 fail-safe evaluation policy of the Scania Component X benchmark. The results demonstrate that optimisation under a business-driven cost structure leads to markedly different operational behaviours than accuracy-oriented evaluation.

Table 4. Comparison of cost-optimised deep learning models under the 5×5 fail-safe evaluation policy on the Scania Component X test set.

Model	Total Cost	Avg. Cost/Vehicle	FN	FP	TP	TN	Maintenance Rate
CNN (cost-optimised)	44,798	9.01	51	2,378	91	2,448	49.7%
Transformer (cost-optimised)	**40,195**	**8.09**	19	3,255	123	1,571	68.0%
TCN (cost-optimised)	48,260	9.71	**0**	4,826	142	0	100.0%

Among the three architectures, the Transformer achieves the lowest total operational cost (40,195), corresponding to an average cost of 8.09 per vehicle, outperforming both the CNN (9.01) and the TCN (9.71). This confirms that economic optimality does not necessarily coincide with maximal safety or minimal maintenance intervention rates.

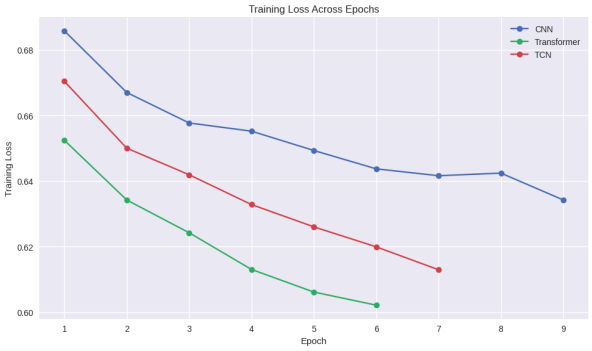


Figure 7. Training loss per epoch for CNN, Transformer, and TCN.

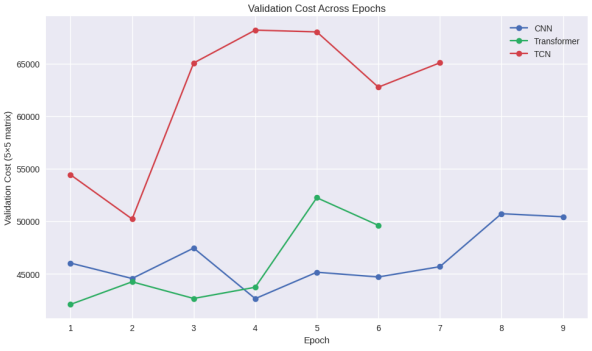


Figure 8. Validation cost per epoch under the 5×5 cost matrix.

The training and validation dynamics of the models are illustrated in Figures 7 and 8. While all architectures exhibit stable convergence during training, the Transformer reaches lower validation cost earlier and maintains a narrower variance across epochs. This behaviour indicates that self-attention mechanisms are particularly effective at aligning temporal representations with economically relevant degradation patterns, rather than simply minimizing classification loss. In contrast, the TCN exhibits pronounced volatility in validation

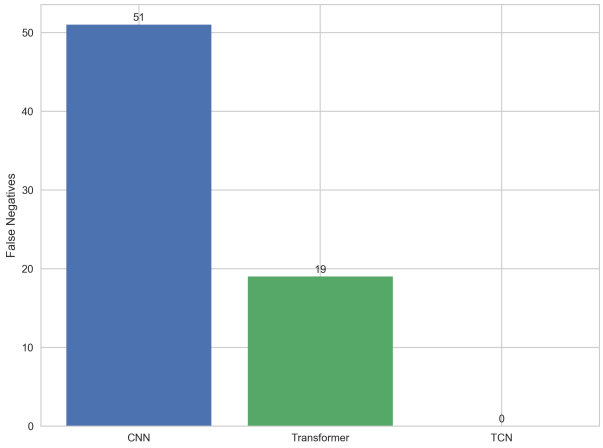
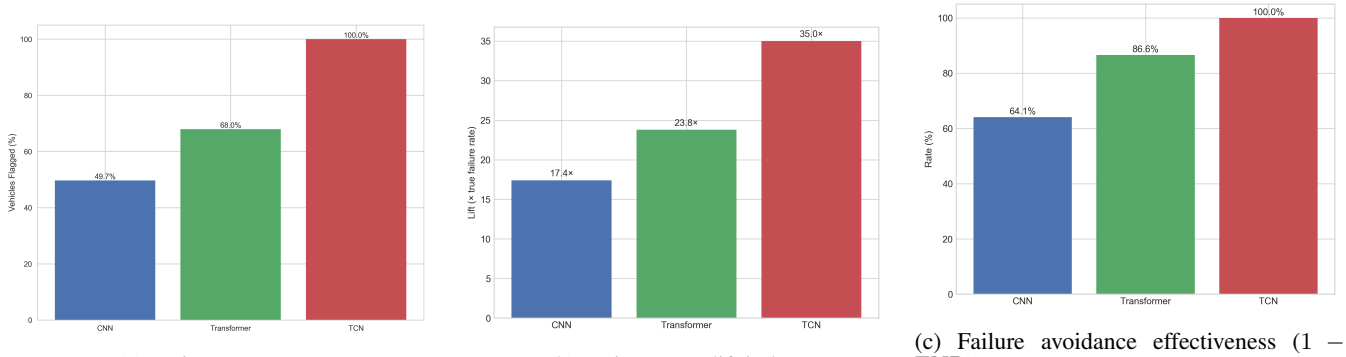


Figure 9. False negatives across the evaluated models.

cost, despite smooth training loss curves, reflecting increased sensitivity to the asymmetric penalty structure of the evaluation matrix.

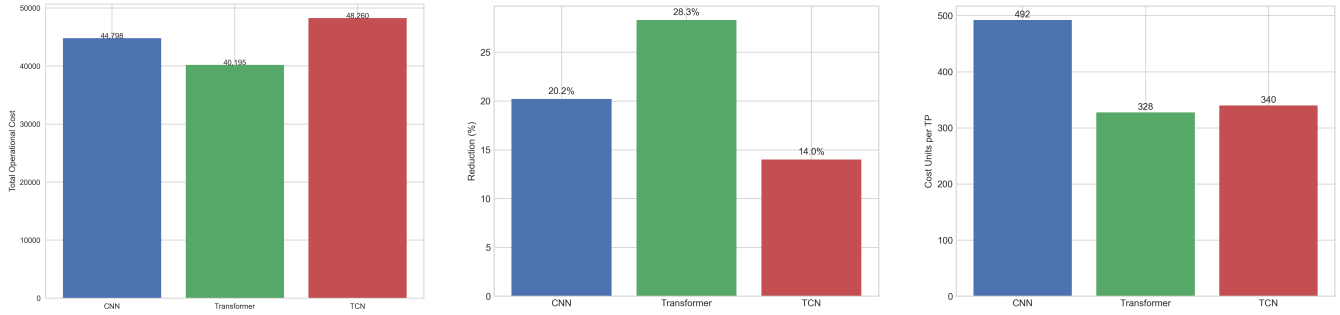
Failure-risk behaviour is further analyzed through the false negative counts shown in Figure 9. The TCN completely eliminates false negatives, achieving perfect failure avoidance. However, this behaviour is achieved by recommending maintenance for the entire fleet, resulting in a 100% maintenance rate (Table 4). The Transformer substantially reduces the number of missed failures compared to the CNN (19 vs. 51), while avoiding the extreme over-maintenance observed for the TCN. This intermediate positioning is particularly important in industrial settings, where eliminating all failures is often economically infeasible due to excessive maintenance workload.

Taken together, Figure 10 reveal how the three architectures navigate the trade-off between failure prevention and operational burden under the asymmetric 5×5 cost matrix. As shown in Figure 10c, the TCN achieves perfect failure avoidance (100%), completely eliminating false negatives, whereas the Transformer and CNN reach 86.6% and 64.1%, respectively. However, this safety guarantee comes at a substantial cost: Figure 10a shows that the TCN flags the entire fleet for preventive maintenance, resulting in a 100% maintenance rate, which in turn produces the highest maintenance lift of $35.0\times$ relative to the true failure rate (Figure 10b). In contrast, the Transformer occupies a more balanced op-



(a) Maintenance rate (b) Maintenance lift index (c) Failure avoidance effectiveness (1 − FNR)

Figure 10. Impact of model choice on maintenance behaviour.



(a) Total operational cost (b) Relative cost reduction (c) Cost per true maintenance action

Figure 11. Economic impact of model choice under the 5×5 cost matrix.

erating regime, combining strong failure avoidance (86.6%) with a lower—though still conservative—maintenance rate of 68.0%, corresponding to a lift of 23.8×. The CNN lies at the opposite end of the spectrum, issuing the fewest maintenance actions (49.7%) but incurring the weakest failure protection, as reflected by its elevated false negative rate. These combined observations indicate that while aggressive maintenance policies guarantee safety, they may impose disproportionate operational workloads, whereas the Transformer achieves the most favourable compromise between failure prevention and maintenance intensity under Scania's cost assumptions.

Economic outcomes of these behaviours are consolidated in Figure 11. As shown in Figure 11a, the Transformer achieves the lowest total operational cost on the test set, despite not achieving perfect failure avoidance. Relative to the all-healthy baseline, the Transformer delivers the largest cost reduction (28.3%; Figure 11b), compared to 20.2% for the CNN and only 14.0% for the TCN. Moreover, Figure 11c shows that the Transformer minimizes the cost per true maintenance action (328), demonstrating superior economic efficiency in converting maintenance interventions into prevented failures.

The qualitative trade-offs summarized in Table 5 further contextualize these quantitative findings. While the TCN provides maximal safety guarantees, its computational burden and excessive maintenance aggressiveness limit its practicality for large-scale fleet deployment. The CNN offers a computationally efficient and stable baseline but fails to sufficiently control failure risk under the given cost structure. The Transformer consistently emerges as the most balanced solution, offering strong failure-risk control, favourable economic efficiency, and robust generalization.

Table 5. Qualitative trade-offs for industrial deployment.

Criterion	**CNN**	**Transformer**	**TCN**
Economic Efficiency	Moderate	**High**	Low
Failure-Risk Control	Moderate	**Strong**	**Perfect**
Maintenance Aggressiveness	Balanced	Conservative	Over-maintaining
Computational Load	Low	Moderate	Highest
Interpretability	Moderate	**High**	Limited
Operational Robustness	Stable	**Strong**	Sensitive
Overall Assessment	Baseline	**Best trade-off**	Safe but heavy

Finally, Table 6 places the proposed models within the Component X literature. The achieved average cost per vehicle is competitive with previously reported classical machine learning, survival analysis, and graph-based approaches, even without contextual feature engineering or model-specific heuristics.

Table 6. Average cost per vehicle compared with published Component X baselines.

Model	Avg. Cost/Vehicle	Reference
Transformer	8.09	This study
TCN	9.71	This study
CNN	9.01	This study
Contextual XGBoost	8.2–9.5	Carpentier et al.
Survival models	9–11	Carpentier et al.
Classical ML (best)	8.3–10.2	Yang and Iqbal
GNN models	9–12	Parton et al.

4. DISCUSSION

The experimental results show how temporal deep learning models behave under the asymmetric cost structure of the Scania Component X benchmark. Because missed failures are much more costly than unnecessary maintenance actions, the models tend to operate in risk-averse settings Kharazian et al. (2025).

For **RQ1**, the models show clear differences in behaviour. The TCN removes all false negatives, but only by recommending maintenance for the entire fleet. The CNN recommends fewer maintenance actions but misses more failures. The Transformer offers a more balanced behaviour by reducing missed failures without the extreme maintenance workload of the TCN.

For **RQ2**, the results show that cost-sensitive optimisation directly affects model decisions. Since false negatives are penalised much more heavily than false positives, the models favour failure detection even when this increases preventive maintenance actions.

For **RQ3**, the Transformer provides the best balance between operational cost, failure avoidance, and maintenance workload. While the TCN maximises safety and the CNN reduces maintenance actions, the Transformer achieves the lowest total operational cost while keeping a more practical operating regime.

5. CONCLUSION

This study presented a cost-sensitive deep learning approach for maintenance decision-making on the Scania Component X dataset. By reformulating the degradation prediction task into a binary risk estimation problem while remaining compatible with the official asymmetric cost matrix, the proposed approach aligns model training and evaluation with real industrial maintenance objectives.

The experimental results demonstrate that cost-sensitive modelling enables data-driven systems to support economically meaningful maintenance decisions in large industrial fleets. In particular, the evaluated temporal architectures show that balancing failure prevention and maintenance workload is essential for minimizing total operational cost.

Future work will explore adaptive decision thresholds, richer temporal representations, and the integration of predictive models with prescriptive maintenance planning strategies.

REFERENCES

Carpentier, L., De Temmerman, A., & Verbeke, M. (2024). Towards contextual, cost-efficient predictive maintenance in heavy-duty trucks. In *International symposium on intelligent data analysis* (pp. 260–267).

Chehade, A., et al. (2022). Conditional gaussian mixture model for warranty claims forecasting. *Reliability Engineering & System Safety*, *218*, 108180. doi: 10.1016/j.ress.2021.108180

Deldari, S., Loke, S. W., & Zaslavsky, A. (2023). Predictive maintenance using deep learning: Review, challenges, and industrial applications. *Sensors*, *23*(13), 6058. doi: 10.3390/s23136058

Ding, B., & Fang, H. (2017). Fault prediction for nonlinear stochastic systems with incipient faults. *ISA Transactions*, *68*, 327–334.

Dintén, R., Zorrilla, M., Veloso, B., & Gama, J. (2025). Building of transformer-based rul predictors supported by explainability techniques: application on real industrial datasets. *Information Fusion*, 103892.

Johnson, J. M., & Khoshgoftaar, T. M. (2021). Thresholding strategies for deep learning with highly imbalanced big data. In *Advances in intelligent systems and computing* (Vol. 1232, pp. 123–136). Springer. doi: 10.1007/978-981-15-6759-9_9

Kharazian, Z., Lindgren, T., Magnússon, S., Steinert, O., & Andersson Reyna, O. (2025). Scania component x dataset: A real-world multivariate time-series dataset for predictive maintenance. *arXiv preprint arXiv:2401.15199*.

Khoshkangini, R., Mashhadi, P. S., Tegnered, D., & Rognvaldsson, T. (2022). Vehicles behavioral prediction using multi-task ensemble learning. *SSRN Electronic Journal*. doi: 10.2139/ssrn.4087639

Khoshkangini, R., Pashami, S., & Nowaczyk, S. (2019). Warranty claim rate prediction using logged vehicle data. In *Progress in artificial intelligence* (Vol. 11804, pp. 663–674). Springer. doi: 10.1007/978-3-030-30241-2_55

Lei, Y., et al. (2020). Applications of machine learning to machine fault diagnosis: A review. *Mechanical Systems and Signal Processing*, *138*, 106587.

Man, J., & Zhou, Q. (2018). Prediction of hard failures using stochastic degradation signals. *Computers & Industrial Engineering*, *125*, 480–489.

Mashhadi, P. S., et al. (2019). Stacked ensemble of recurrent neural networks for rul prediction. *Applied Sciences*.

Parton, M., Fois, A., Vegliò, M., Metta, C., & Gregnanin, M. (2024). Predicting the failure of component x in the scania dataset with graph neural networks. In *International symposium on intelligent data analysis* (pp. 251–259).

Rengasamy, D., Jafari, M., & Rothwell, B. (2020). Deep learning with dynamically weighted loss function for prognostics and health management. *Sensors*, *20*(3), 723. doi: 10.3390/s20030723

Shah, S., Wilder, B., Perrault, A., & Tambe, M. (2024). Leaving the nest: Going beyond local loss functions for predict-then-optimize. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 38, pp. 14902–14909).

Xu, D., & Zhang, J. (2026). Risk-aware decision-making for predictive maintenance using distributional reinforcement learning. *International Journal of Systems Assurance Engineering and Management*. doi: 10.1007/s40747-025-02127-w

Yang, Y., & Iqbal, M. Z. (2025). Cost-optimised machine learning model comparison for predictive maintenance. *Electronics*, *14*(12), 2497.

Zhang, Q., Wang, J., & Li, T. (2023). A comprehensive review of cost-sensitive learning in predictive maintenance. *Reliability Engineering & System Safety*, *238*, 109444. doi: 10.1016/j.ress.2023.109444