

Predictive Quality-Based Estimation of Lifetime Testing in Durable Home Appliances Using Unsupervised Deep Learning

Yigit Elarslan, Ceren Eda Cicek and Efe Ozkaynak

Undergraduate, Department of Industrial Engineering

Eskisehir Technical University, Eskisehir, Türkiye

elarslanyigit@hotmail.com, edacerencicek@gmail.com, kardelenefezade1@gmail.com

Gulcin Dinc Yalcin

Associate Professor Doctor, Department of Industrial Engineering

Eskisehir Technical University, Eskisehir, Türkiye

gdinc@eskisehir.edu.tr

Eren Guvenilir

Department of Physics

Eskisehir Osmangazi University, Eskisehir, Türkiye

Haier Europe, Eskisehir, Türkiye

eren.guvenilir@haier-europe.com

Abstract

Lifetime testing in the durable home appliances sector is a time and resource-intensive validation process, especially for tumble dryers that require 1600 operating cycles to represent the estimated product lifetime. In practice, intermediate decisions are often made around the 320th cycle to accelerate product development; however, these decisions largely depend on expert judgment and may fail to detect failures that occur in later cycles. This study proposes a data-driven Predictive Quality (PdQ) framework to support early and more reliable evaluation of tumble dryer lifetime tests. Due to the limited availability of faulty samples and the difficulty of identifying exact fault onset points in long sensor sequences, an unsupervised deep learning (DL) approach was adopted. An LSTM-based autoencoder (AE) was trained using healthy C-class machine data to learn normal operating behavior from multivariate time-series signals, including power consumption, temperature variables, and control signals. The trained model was then evaluated on E-class machine data containing both healthy and faulty test cases. Reconstruction error was used as the anomaly score, and a multi-level thresholding structure was applied to classify test behavior into normal, warning, and critical risk levels. The model achieved an F1-score of 82.35%. The proposed workflow includes data cleaning, preprocessing, model training, anomaly score generation and decision support interpretation. The results indicate that the proposed approach has the potential to generate early risk indicators, transform raw test data into actionable insights, support operator decisions, and improve the efficiency of lifetime testing processes.

Keywords

Predictive Quality, Deep Learning, Decision Support System, Anomaly Detection, Lifetime Testing.

1. Introduction

Increasing global competition in the durable home appliances sector has made the improvement and acceleration of R&D and quality control processes a strategic requirement. In particular, lifetime tests conducted to verify the reliability of tumble dryers require 1600 operating cycles, which correspond to the estimated 10-year lifetime of the product. These long-duration tests result in significant time, energy, labor, and laboratory capacity costs. In industrial practice, however, companies often need to make earlier decisions in order to accelerate R&D and product development processes. In the company where this study was conducted, an intermediate quality decision is made after the 320th cycle, which corresponds to the 2-year warranty point. At this stage, the decision process is mainly based on the heuristic judgment and experience of technical staff. Although this approach provides practical support for early evaluation, it also carries the risk of missing potential failures that may occur after the 320th cycle. As a result, projects may need to return to earlier design or validation stages to correct late-detected failures, which can slow down R&D activities and extend product development cycles.

The company conducts lifetime tests on tumble dryers belonging to different energy efficiency classes. In line with the European Union energy labelling framework for household tumble dryers, this study uses the energy class distinction as a reference point for model validation (European Union, 2023). The main motivation is to transform operational data collected from the existing data acquisition (DAQ) infrastructure into meaningful indicators that can support test monitoring and decision-making. Due to the nature of industrial lifetime testing, faulty samples are rare, while healthy operating data are relatively abundant. This data imbalance makes conventional supervised learning approaches less suitable, as they generally require sufficient and balanced datasets. In addition, manually labeling the exact start time of a fault in long sensor sequences is costly, time-consuming, and prone to expert-dependent uncertainty. Therefore, this study adopts an unsupervised DL approach based on AE architecture. The proposed model learns the normal operating characteristics of the system using healthy lifetime test data. Deviations from this learned behavior are quantified through reconstruction error and interpreted as anomaly scores. In this way, the model can detect not only known fault patterns but also previously unseen abnormal behaviors that differ from the healthy reference profile. The main contribution of this study is not the proposal of a new autoencoder architecture, but the practical adaptation of an unsupervised reconstruction-based anomaly detection approach to a predictive quality decision support problem in long-duration tumble dryer lifetime testing. In this context, the study contributes to the literature and industrial practice in five main aspects. First, it addresses lifetime testing as a predictive quality problem rather than a conventional predictive maintenance task, focusing on early quality risk assessment before the completion of the full test campaign. Second, it develops a data strategy that explicitly considers energy-class-dependent operating behavior as a potential source of domain shift in tumble dryer data. Third, it transforms LSTM-autoencoder reconstruction errors into operator-friendly decision levels through a multi-level interpretation structure. Fourth, it improves the interpretability of the anomaly detection output by analyzing feature-based reconstruction error contributions and linking them with physically meaningful variables. Finally, it provides a preliminary industrial validation using real company test data, where the proposed framework is evaluated at the test level under limited but realistic healthy and faulty sample conditions. Therefore, the study positions the LSTM-autoencoder not as a standalone fault classifier, but as a decision-support component that can generate early risk indicators and support data-driven quality evaluation in durable home appliance lifetime testing.

1.1. Objectives

The main objective of this study is to develop a data-driven approach that can support the early evaluation of tumble dryer lifetime tests without compromising quality assurance requirements. To achieve this objective, the study first aims to establish a structured data strategy for the measurements collected through the existing DAQ infrastructure. The second objective is to develop a Long Short-Term Memory (LSTM) based AE model that learns healthy operating behavior from multivariate time-series data. The final objective is to evaluate the proposed approach on healthy and faulty company test data and examine whether the generated anomaly scores can provide meaningful support for test-level evaluation. In addition to numerical performance metrics, feature-based error analysis is used to improve interpretability and to help quality personnel understand which variables contribute most to abnormal behavior. In this way, the study aims to establish a foundation for reducing unnecessary test cycles, minimizing energy-related costs, and supporting proactive quality assurance in lifetime testing processes.

2. Literature Review

Predictive Maintenance (PdM) and PdQ approaches have gained importance with Industry 4.0 due to the increasing availability of sensor data in industrial systems. PdM studies generally focus on fault prediction, anomaly detection,

remaining useful life estimation, fault classification, time-series forecasting, and condition monitoring (Dalzochio, et al., 2020). Although these approaches were first developed mainly for maintenance applications, similar data-driven methods can also be adapted to quality control and R&D validation processes. In this study, PdM-based anomaly detection is adapted to the lifetime testing of tumble dryers as a PdQ decision support problem. While PdM focuses on predicting when equipment will fail and on scheduling maintenance actions to maximize uptime, PdQ targets the quality of the product or process output rather than the health of the equipment itself. In the context of tumble dryer lifetime testing, PdQ refers to the early identification of test records whose operating behavior is inconsistent with a product expected to pass quality validation, enabling accelerated qualification without completing the full 1600-cycle protocol. This distinction is important: a PdM framework for this setting would predict when the test bench machinery might fail, whereas the proposed PdQ framework predicts whether the product under test exhibits behavior inconsistent with a product that would pass—regardless of whether any equipment failure has occurred. Recent PdQ literature further supports this perspective. Tercan and Meisen (2022) reviewed machine learning and deep learning-based predictive quality studies in manufacturing and showed that PdQ applications commonly aim to predict product or process quality before final inspection by using production data, process variables and sensor-based measurements. Similarly, Kausik et al. (2025) reviewed machine learning algorithms for manufacturing quality assurance and emphasized that model selection in industrial quality applications should consider not only accuracy but also scalability, computational efficiency, real-time applicability, and suitable performance metrics. These studies indicate that predictive quality is not limited to final defect classification; it also includes earlier quality risk estimation and data-driven support for operational decision-making. In this context, the present study contributes to PdQ literature by focusing on long-duration R&D lifetime tests, where the aim is to generate early quality risk indicators before the full test campaign is completed. Several studies emphasize that industrial machine learning applications face challenges such as scalability, dynamic operating conditions, high-dimensional sensor data, and limited generalization across different machines (Dalzochio, et al., 2020). Carbery et al. (2018) highlighted the importance of preprocessing, feature selection, and hybrid modeling strategies for large-scale manufacturing data, while Li et al. (2017) pointed out that models trained on data from a single machine may not generalize well to other devices or operating conditions. These findings support the need for a robust data strategy before model development, especially in manufacturing environments with multiple product types and operating profiles. This issue is particularly relevant for tumble dryer lifetime testing because different energy classes may naturally differ in power consumption, temperature behavior, and control dynamics. Another major challenge in industrial quality and fault detection problems is data imbalance. In real production and test environments, healthy operating data are usually abundant, whereas faulty data are rare and difficult to label. Krawczyk (2016) stated that conventional supervised classifiers may become biased toward the majority class in imbalanced datasets and fail to detect rare but critical faulty cases. In addition, labelling the exact starting point of a failure in long lifetime test sequences is costly and requires expert knowledge. Therefore, unsupervised or semi-supervised anomaly detection approaches are more suitable for this study. AE-based DL models provide an effective solution to this problem because they can be trained only on healthy data and detect deviations from normal behavior through reconstruction error. Zhao et al. (2019) emphasized that deep learning models, especially autoencoders, can automatically learn hierarchical features from raw sensor data and reduce the need for manual feature engineering.

Aslam et al. (2024) also used reconstruction error as an anomaly score after training an AE on normal operating data. This principle is directly aligned with the aim of the present study, where the model learns healthy tumble dryer behavior and identifies abnormal test behavior through reconstruction error. Since lifetime test data are multivariate time series, temporal dependencies between consecutive sensor readings are critical. Recent multivariate time-series anomaly detection literature has expanded from reconstruction-based recurrent models toward attention-based and transformer-based architectures. Wang et al. (2025) provided a structured survey of deep anomaly detection methods for multivariate time series and classified recent studies according to learning paradigms and model architectures. TranAD, proposed by Tuli et al. (2022), is an example of a modern transformer-based anomaly detection model that uses attention-based sequence encoders, self-conditioning, and adversarial training to improve anomaly detection and diagnosis in multivariate time-series data. These studies show that the field is moving toward more advanced temporal representation learning methods. However, such models may require larger datasets, more careful tuning, and higher deployment complexity in small industrial validation settings. Different AE variants have been proposed in the literature, including sparse, denoising, contractive, convolutional, and recurrent autoencoders (Zhang et al. (2018); Li et al. (2023)). Among these alternatives, LSTM-based autoencoders are particularly suitable for sequential industrial data because they can model temporal dependencies while preserving a reconstruction-based anomaly detection structure. Malhotra et al. (2016) showed that LSTM encoder-decoder models can capture temporal dependencies in multi-sensor time-series anomaly detection problems. Similarly, Ahmad et al. (2020) applied LSTM autoencoders to

rotating machine condition monitoring and reported strong anomaly detection performance using the NASA IMS Bearing dataset vibration data (Lee et al. 2007). More recently, Lee et al. (2024) used an LSTM-autoencoder-based unsupervised anomaly detection approach on wind turbine vibration data, demonstrating that LSTM-AE structures remain relevant for learning normal patterns and detecting deviations in industrial time-series systems. Although the present study does not use vibration signals, these studies demonstrate that LSTM-AE architectures are effective for learning sequential dependencies and detecting abnormal behavior from sensor-based operational data. Therefore, LSTM-AE was selected in this study as a practical and interpretable reconstruction-based architecture suitable for limited faulty data, sequential sensor patterns, and DSS-oriented output interpretation. For model evaluation, the reconstruction error was compared with predefined threshold levels to classify the system state as normal, warning, or critical. This structure was adapted from the traffic-light-based anomaly interpretation approach proposed by Givnan et al. (2022) for industrial motor monitoring.

In the present study, this logic was used to transform reconstruction-error outputs into operator-friendly decision levels within the PdQ decision support framework. Performance is evaluated using a confusion matrix and metrics such as accuracy, precision, recall, F1-score, false positive rate, false negative rate, and Matthews Correlation Coefficient (Chicco & Jurman, 2020). These metrics are preferred because accuracy alone may be misleading in imbalanced industrial datasets (Krawczyk 2016; Aslam et al. 2024; Givnan et al. 2022). Existing predictive maintenance studies have widely used reconstruction-based anomaly detection for machine health monitoring; however, these studies mainly focus on maintenance-oriented fault detection or remaining useful life estimation. In contrast, predictive quality applications in manufacturing mostly address production-stage quality prediction or final defect estimation. Long-duration R&D lifetime tests, where the objective is to support early quality evaluation before the completion of the full test campaign, remain less explored. This setting introduces additional challenges, including scarce faulty samples, uncertain fault-onset labels, test-level rather than window-level ground truth, and operating-profile differences across energy classes. Therefore, there is a need for a practical PdQ decision-support framework that can transform reconstruction-error patterns into interpretable risk levels for lifetime test monitoring rather than acting as a fully autonomous fault classifier. Overall, LSTM-AE models are well-suited for detecting abnormal behavior in multivariate industrial time-series data when faulty labels are limited. However, existing studies predominantly address maintenance-oriented fault detection for rotating machinery, while long-duration R&D lifetime tests for consumer appliances -where the objective is early quality risk assessment rather than equipment maintenance scheduling- remain underexplored. In this setting, additional challenges arise ground-truth labels are available only at the test-record level, the number of faulty records is small, and the model must generalize across energy-class profiles not seen during training. The present study targets this specific gap by adapting an LSTM-AE reconstruction-error framework to tumble dryer lifetime testing and embedding its output into a PdQ decision support structure with operator-interpretable risk levels.

3. Methods

The overall process consists of five main stages: current system analysis, data preparation, unsupervised model training, reconstruction-error-based anomaly scoring, and decision support interpretation. The general workflow of the proposed methodology is shown in Figure 1. The process starts with the analysis of the current lifetime testing system and the definition of the PdQ problem and objectives. Then, multivariate operational data are analyzed, cleaned, and preprocessed. After preprocessing, the LSTM-based AE model is trained using healthy operating data. If the training performance is not acceptable, model parameters are tuned and the training process is repeated. Once the model reaches an acceptable training behavior, reconstruction errors are generated and risk levels are applied. Finally, the outputs are evaluated in terms of their suitability for decision support. If the outputs are not meaningful for operational interpretation, the process returns to the data preparation stage for further refinement. If the outputs are valid, they are transformed into decision support outputs for prototype deployment.

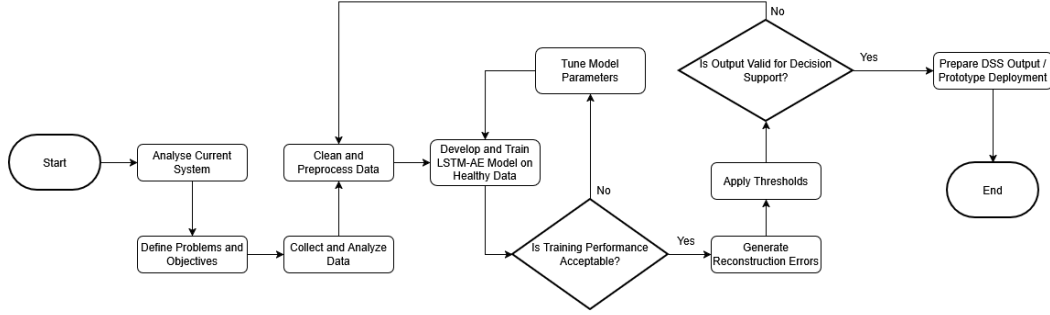


Figure 1. Flowchart of the proposed Predictive Quality decision support methodology.

The dataset consists of multivariate time-series signals representing power, temperature, and control behavior. In preprocessing stage, inconsistent observations, missing values, and signal-level irregularities were handled before model training. The preprocessed data were then structured into fixed-length sequential windows. This window-based structure enables the model to learn not only instantaneous sensor values but also temporal dependencies among consecutive measurements. In the modelling stage, an LSTM-AE architecture was used. Autoencoders are neural network models that learn a compressed latent representation of the input and reconstruct the original data from this representation. In this study, the AE principle is used for anomaly detection. The model is trained only on healthy operating data; therefore, healthy patterns are expected to be reconstructed with relatively low errors. In contrast, faulty or abnormal operating patterns are expected to generate higher reconstruction errors because they deviate from the learned healthy reference behavior. The recurrent structure of the model is based on LSTM cells. LSTM networks are suitable for sequential industrial data because they can model temporal dependencies and preserve information from previous time steps. This property is important for lifetime test data, where current machine behavior may depend on previous patterns. The architecture used in this study is presented in Figure 2.

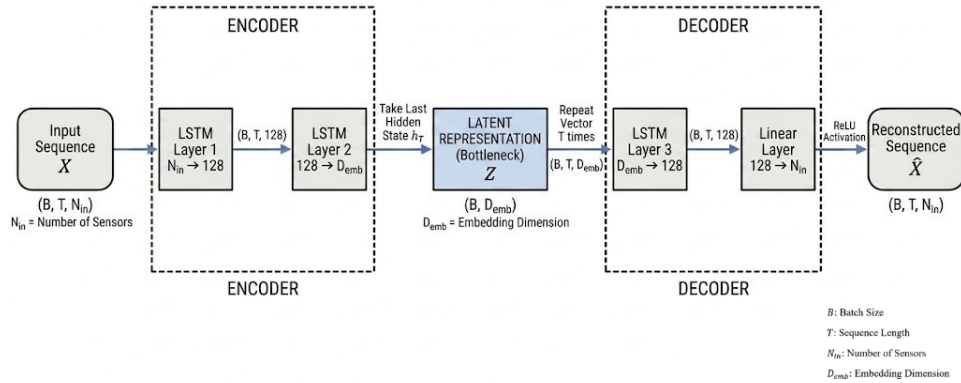


Figure 2. LSTM-based autoencoder architecture used for reconstruction-based anomaly detection.

The input sequence is represented as:

$$X_i \in \mathbb{R}^{T \times P}$$

where X_i denotes the i^{th} input window, T is the sequence length, and P is the number of input variables. In implementation, the input tensor is processed in batch form as:

$$X \in \mathbb{R}^{B \times T \times N_{in}}$$

where B is the batch size and N_{in} is the number of input sensor variables. In the encoder section, the first LSTM layer maps the input sequence into a 128-dimensional hidden representation. The second LSTM layer compresses this

representation into a lower-dimensional embedding space with dimension D_{emb} . The final hidden state of the encoder is used as the latent representation:

$$z_i = Enc_{\theta}(X_i)$$

where z_i summarizes the temporal behavior of the input window. In the decoder section, the latent representation is repeated across the sequence length and passed through an LSTM layer to reconstruct the temporal structure of the original input. A final linear layer maps the decoder output back to the original input dimension. The reconstructed sequence is obtained as

$$\hat{X}_i = Dec_{\theta}(z_i) = f_{\theta}(X_i)$$

where $\hat{X}_i$ denotes the reconstructed version of the original input window and θ represents the trainable model parameters. The encoder compresses the multivariate input sequence into a lower-dimensional latent representation, while the decoder reconstructs the original sequence from this representation. The reconstruction error between the input sequence and the reconstructed output is used as the anomaly score. The model parameters are optimized by minimizing the reconstruction loss between the original input sequence and the output sequence. For each observation, the reconstruction error is defined as:

$$e_{i,t,j} = x_{i,t,j} - \hat{x}_{i,t,j}$$

where $x_{i,t,j}$ is the original value of the j^{th} variable at time step t in window i , and $\hat{x}_{i,t,j}$ is the corresponding reconstructed value. To reduce the excessive influence of isolated large deviations during training, the reconstruction objective was formulated using the SmoothL1 loss. This loss behaves quadratically for small reconstruction errors and linearly for larger deviations:

$$\ell_{\beta}(e) = \begin{cases} \frac{1}{2} \frac{e^2}{\beta}, & \text{if } |e| < \beta \\ |e| - \frac{1}{2}\beta, & \text{otherwise} \end{cases}$$

Accordingly, the overall training objective is expressed as:

$$\mathcal{L}_{SmoothL1}(\theta) = \frac{1}{NTP} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^P \ell_{\beta}(x_{i,t,j} - \hat{x}_{i,t,j})$$

where N is the number of training windows, T is the sequence length, and P is the number of input variables. SmoothL1 loss is suitable for industrial lifetime test data because sensor noise, transient operating changes, and imperfectly filtered observations may exist even in data accepted as healthy. Therefore, the loss function preserves sensitivity to normal reconstruction differences while preventing a small number of extreme windows from dominating the optimization process. For each test window, the anomaly score is calculated as the average squared reconstruction error across all time steps and variables:

$$s_i = \frac{1}{TP} \sum_{t=1}^T \sum_{j=1}^P (x_{i,t,j} - \hat{x}_{i,t,j})^2$$

A low anomaly score indicates that the observed window is consistent with the learned healthy operating behavior. A high anomaly score indicates that the temporal pattern of the test window deviates from the healthy reference profile learned by the model. To make the reconstruction error output suitable for decision support, a multi-level thresholding strategy was applied. Percentile-based thresholds were derived from the reconstruction error distribution of healthy validation windows. This prevents information leakage because the threshold values are not calculated from test data. Let the set of validation anomaly scores be defined as:

$$S_{val} = \{s_1, s_2 \dots s_m\}$$

The warning and critical thresholds are calculated as percentile values of the healthy validation score distribution:

$$\begin{aligned}\tau_w &= Q_{q_2}(S_{val}) \\ \tau_c &= Q_{q_3}(S_{val})\end{aligned}$$

where $Q_q(\cdot)$ represents the q^{th} percentile of the validation score distribution, τ_w is the warning threshold, and τ_c is the critical threshold. The operational risk level of a test window is then defined as:

$$Risk(s_i) = \begin{cases} Normal, & s_i \leq \tau_w \\ Warning, & \tau_w < s_i \leq \tau_c \\ Critical, & s_i > \tau_c \end{cases}$$

This multi-level structure is more appropriate for industrial decision support than a single hard threshold. It allows the model output to distinguish between acceptable healthy variation, suspicious deviation, and critical abnormal behavior. Therefore, decision makers can interpret the model output gradually rather than relying on an immediate binary healthy/faulty classification. The selected architecture size was considered sufficient for the available dataset and the targeted performance level. However, the number of LSTM layers, hidden units, embedding dimension, and sequence length are not fixed design choices for all industrial cases. These parameters may vary depending on the number of input variables, the amount of available training data, the complexity of the machine behavior, and the expected anomaly sensitivity. Therefore, for larger datasets or different product groups, the model architecture can be re-tuned to improve generalization and detection performance. Finally, model performance was evaluated at the test level using the confusion matrix and related classification metrics. In addition to numerical performance metrics, feature-based reconstruction error analysis was performed to identify which variables contributed most to the anomaly score. This step supports interpretability by linking the model output with physically meaningful machine behavior.

4. Data Collection and Experimental Setup

The data used in this study were collected from tumble dryer lifetime tests conducted in the company's R&D laboratory. The available data were obtained from the existing DAQ infrastructure used in the laboratory. The main variables considered in this study include power consumption, temperature measurements, and control signals related to the thermal and operational behavior of the machine. In particular, hot air temperature from front duct (represented as T2) and compressor outlet (represented as T3) temperature signals, power values, and compressor on/off signal (represented as binary HEATER CMD) were selected because they provide information about heating behavior, compressor-related thermal response, and control logic during dryer operation. Since the collected measurements are time-dependent, the data were organized as multivariate time series. To improve the reproducibility and transparency of the experimental design, the dataset structure and model configuration are summarized in Table 1. The training data consisted only of healthy C-class machine records, while the external evaluation was performed on 8 kg E-class machine records including both healthy and unhealthy test cases. Preliminary data analysis showed that different energy classes may exhibit distinct power consumption, temperature response, and control behavior patterns. Therefore, energy class was accepted as the main segmentation axis in the data strategy. This separation was necessary to prevent the model from interpreting natural operating differences between energy classes as anomalies during training. However, due to the available industrial data structure, the initial experimental evaluation was designed as a preliminary cross-class robustness assessment. Healthy C-class data were used to learn the reference operating behaviour, while 8 kg E-class data, including both healthy and faulty test cases, were used during inference to examine whether the model could still generate meaningful anomaly indicators under a different energy-class operating profile. The faulty test group was formed based on information provided by the industrial advisor. Therefore, the reported performance should be interpreted as reflecting both anomaly detection capability and preliminary cross-class robustness, rather than as a pure same-domain fault detection benchmark. Before model training, raw data were cleaned, temporally aligned, and transformed into a consistent multivariate time-series format. The raw records were processed through source-specific preprocessing steps, including timestamp parsing, numerical type conversion, missing-value handling, and temporal alignment. The time-series records were resampled to a 1-second frequency in order to obtain a consistent sampling structure across the selected variables. Missing or inconsistent numerical observations were handled during preprocessing, and the selected variables were standardized using a StandardScaler fitted only on the healthy training data. The same scaler parameters were then applied to validation and inference data

to avoid information leakage. The healthy reference dataset was divided into training and validation subsets at the cycle level rather than at the individual window level. In the final experimental run, 32 healthy C-class reference records were used for training, and 9 healthy C-class reference records were used for validation. These records were collected from 7 C-class tumble dryer machines, providing variation in test history and machine-level operating characteristics across the training data. These records produced 15,424 training windows and 4,186 validation windows, respectively. The external test-level evaluation was performed on 17 labelled E-class records, including 7 healthy records and 10 unhealthy records. The unhealthy group was defined based on field information provided by the industrial advisor, and the known fault mechanism was associated with a compressor-related problem. Since the exact window-level fault onset was not available, the labels were used at the test-record level. The time-series data were divided into fixed-length sequential windows with a sequence length of 20-time steps.

Since the data were resampled to a 1-second frequency, each input window represented approximately 20 seconds of machine behavior. Non-overlapping windows were used in the current implementation; therefore, the window stride was equal to the window length, corresponding to a stride of 20-time steps and an overlap ratio of 0%. Remaining rows that did not form a complete final window were discarded. This windowing strategy enabled the LSTM-AE to learn short-term temporal dependencies among consecutive measurements rather than evaluating each observation independently. Since multiple windows can belong to the same physical test record, the final performance evaluation was conducted at the test-record level rather than at the window level. This choice prevents the performance metrics from being artificially inflated by highly correlated window-level samples. The final model configuration was selected based on preliminary experiments in which different hidden layer sizes, embedding dimensions, dropout rates, batch sizes, and learning rates were tested. In the encoder, the input sequence was mapped into 128 and 64 hidden units and then compressed into a 32-dimensional latent representation. The decoder followed a symmetric structure to reconstruct the original multivariate sequence. Adam optimizer was used with a learning rate of 0.0003, batch size of 64, and dropout rate of 0.05. Threshold values were calculated from the reconstruction-error distribution of healthy C-class reference windows. The E-class test data were not used during threshold selection; therefore, the thresholding process did not include test-set information. A percentile-based strategy was preferred because the absolute scale of reconstruction error may vary depending on preprocessing, selected variables, model configuration, and data distribution. Therefore, thresholds were defined relative to the healthy reference distribution rather than as fixed universal error values. In the final configuration, the 95th percentile was used as the warning-level boundary, while the 99th percentile was used as the critical-level boundary. The 99th percentile was selected because it represents the upper tail of the healthy reference error distribution and provides a conservative decision boundary that limits false alarms from transient operating fluctuations in healthy records. A threshold sensitivity check shown in Table 1 confirmed this choice: lower percentiles raised the critical-window ratio in healthy records, while higher percentiles reduced sensitivity to faulty patterns; the 99th percentile offered the best balance between sensitivity and false-alarm control. Since the model generates anomaly scores at the window level, the final numerical evaluation was performed at the test-record level: a test record was classified as suspicious or unhealthy when its reconstruction-error pattern showed critical-level deviations according to the multi-level threshold structure.

Table 1. Threshold sensitivity check based on critical-window ratios

Critical percentile	Critical threshold	Healthy critical windows	Faulty critical windows	Interpretation
95.0	0.022178	112 / 3,426 (3.27%)	9,650 / 10,509 (91.83%)	Highly sensitive, but higher false-alarm risk
97.5	0.271110	22 / 3,426 (0.64%)	7,334 / 10,509 (69.79%)	More conservative, still sensitive
99.0	2.131119	2 / 3,426 (0.06%)	6,306 / 10,509 (60.01%)	Selected operating point
99.5	2.717004	1 / 3,426 (0.03%)	5,507 / 10,509 (52.40%)	Very conservative, lower sensitivity

5. Results and Discussion

5.1 Numerical Results

The numerical evaluation was conducted at the test level using healthy and unhealthy small test groups. The healthy test set included 7 cycles, while the faulty test set included 10 cycles. According to the information provided by the

industrial advisor, a compressor-related fault occurred at around test 800th cycle, and the following 10 tests were considered unhealthy since the machine did not return to its normal operating behavior. At the test level, a sample was classified as suspicious or unhealthy if the reconstruction-error-based anomaly score exceeded the defined multi-level threshold structure and enters the red area. As shown in Table 2, the model produced promising results; however, due to the limited test-level dataset, its reliability should be interpreted cautiously at this stage.

Table 2. Confusion matrix

Actual Condition	Predicted Healthy	Predicted Unhealthy
Healthy	7	0
Unhealthy	3	7

To contextualize the proposed framework, simpler anomaly detection baselines, including PCA-based anomaly detection, Isolation Forest, and One-Class SVM, were also examined. Under the current limited test-level evaluation setting, these methods produced comparable performance to the proposed LSTM-AE. This result should be interpreted cautiously, since the evaluation set contains only 17 labelled test records and different models may obtain identical metrics when they classify the same records correctly or incorrectly. Therefore, the current baseline comparison does not provide sufficient evidence to claim clear performance superiority for the LSTM-AE. Nevertheless, the LSTM-AE was retained as the main framework because it explicitly models sequential dependencies in multivariate lifetime test data, while the tested baselines mainly treat window-level observations as independent feature vectors. Thus, the contribution of the proposed approach lies not in outperforming all baselines on a small test set, but in providing a temporally consistent and interpretable reconstruction-error framework for PdQ-oriented decision support. The resulting performance metrics are presented in Table 3. The model achieved an F1-score of 82.35%. The Matthews Correlation Coefficient was calculated as 0.70, indicating a strong positive association between the model predictions and the actual test conditions. Since the evaluation was performed on only 17 labelled test records, the reported metrics should be interpreted as preliminary indicators of performance rather than definitive estimates of real-world generalization capability.

Table 3. Performance metrics

Metric	Value
F1-Score	82.35%
False Positive Rate	0%
False Negative Rate	30%
Matthews Correlation Coefficient	0.70

5.2 Graphical Results

The graphical results support the numerical findings by showing the reconstruction error behavior of healthy and faulty tests. Figure 3 presents the score overview for the healthy test set. The reconstruction error values remain within the normal operating region, and no warning or critical alarm is observed.

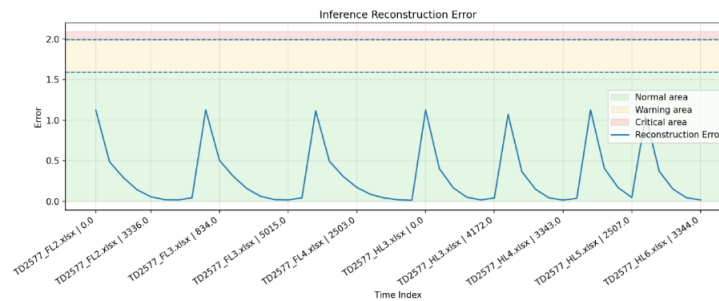


Figure 3. Reconstruction error score overview for healthy test data

Figure 4 presents the score overview for the faulty test group. Several tests exceed the warning or critical anomaly regions, indicating that the model can detect deviations from the healthy reference profile. In particular, test 807 shows the strongest abnormal behavior, with a high number of alarm windows and a clear separation from the normal region. Tests 802, 804, and 810 also present noticeable anomaly patterns. However, tests 801, 805, and 806 remain mostly within the normal region, which explains the false negative cases reported in the numerical results. This result suggests that the current thresholding strategy may not be sufficiently sensitive to all faulty operating patterns. In other words, some unhealthy tests may produce reconstruction errors that are higher than typical healthy behavior but remain below the selected alarm thresholds. Therefore, the missed detections can be interpreted partly as a threshold calibration problem rather than solely as a model learning limitation. From a quality assurance perspective, these false negative cases are particularly important because unhealthy tests may be incorrectly interpreted as normal operating behavior. Consequently, the proposed framework should be considered as a decision-support tool rather than a fully autonomous decision mechanism. To reduce missed detections, future work should focus on threshold refinement, larger validation datasets, and more robust decision logic incorporating consecutive-window behavior and operating-phase-specific anomaly criteria.

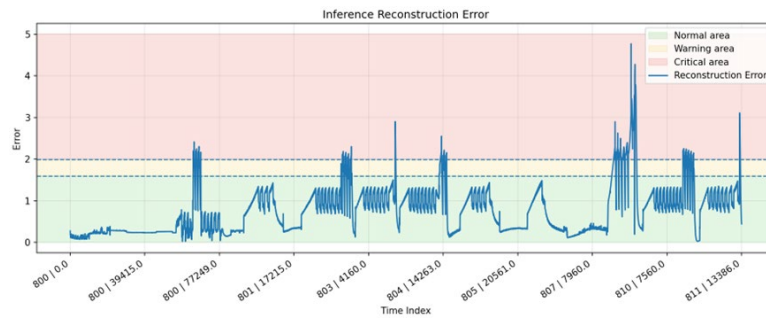


Figure 4. Reconstruction error score overview for faulty test data

Feature-based reconstruction error analysis was also performed to improve interpretability. Figure 5 shows that the highest error contribution in the faulty test group is associated with the HEATER CMD variable, followed by the T2 and T3 temperature variables. Power has a lower average error contribution compared with the other variables. This finding indicates that the detected anomalies are not only related to total power consumption but also to the relationship between control behavior and thermal response.



Figure 5. Feature-based reconstruction error distribution for faulty test data

This result is consistent with the information provided by the industrial advisor. The high error contribution of HEATER CMD supports this information, since this signal is associated with the control behavior of the compressor. Similarly, the T3 error contribution is meaningful because T3 reflects compressor-related temperature behavior. Therefore, the feature-based error analysis strengthens the physical interpretability of the model output.

5.3 Proposed Improvements

Although the obtained results are promising, several improvements are required to increase the model's reliability and quality decision performance. First, the healthy validation set used in this study was limited. Therefore, the false

positive rate of 0% should be interpreted carefully. To validate the model more strongly, the dataset should be expanded with longer and more diverse healthy test records collected from different machines, operating conditions, and test cycles. Second, the thresholding strategy should be recalibrated after increasing the amount and diversity of validation data. Since the current thresholds are derived from a limited healthy reference distribution, they may not fully represent the natural variability of the machine under different test conditions. A larger validation set would allow more reliable estimation of normal, warning, and critical regions. In this way, the model could become more sensitive to faulty behavior while still controlling the false alarm risk. Third, the number of input variables can be increased to represent the machine behavior more comprehensively. The current model uses power, temperature, and control signals; however, additional sensor channels, derivative-based features, drift indicators, and cycle-level trend variables may help capture fault-related behavior that is not sufficiently visible in the current feature set. This improvement is especially important for reducing false negative cases, where faulty tests remain within the learned healthy variation range. Finally, monitoring shifting in reconstruction error trends across successive test cycles may provide a degradation indicator, enabling the framework to evolve toward broader condition monitoring and remaining useful life analysis.

6. Conclusion

This study demonstrated that an unsupervised LSTM-based autoencoder can be used to generate meaningful anomaly indicators for the early evaluation of tumble dryer lifetime tests. The model achieved an F1-score of 82.35%, indicating that reconstruction-error-based anomaly scores can provide meaningful information for distinguishing healthy and unhealthy test behaviour. However, the observed recall of 70% and false negative rate of 30% demonstrate that some unhealthy tests remain undetected. Consequently, the proposed framework should not yet be interpreted as a fully automated acceptance or rejection mechanism. Instead, it should be viewed as a predictive quality decision-support tool that assists engineers by providing early risk indicators and interpretable anomaly information. Importantly, the false positive rate of 0% across all healthy test records indicates that the model does not generate unnecessary alarms on normal operating behavior, which is a critical property for maintaining operator trust in industrial decision support applications. As a preliminary robustness indicator, the threshold sensitivity results show that across the four tested percentile levels, the critical-window ratio in healthy records remained below 3.27% even at the most sensitive threshold, while the ratio in faulty records consistently exceeded 52% at the most conservative threshold. This stable separation across threshold variants suggests that the model's discriminative behavior is not critically dependent on the precise threshold choice within the tested range. Furthermore, the model's ability to identify anomalous windows well before the end of a test record demonstrates a practically relevant early warning capability that supports proactive quality evaluation rather than post-hoc fault analysis. Feature-based analysis confirmed that HEATER CMD and compressor temperature variables drove the anomaly scores, consistent with the known fault mechanism. Overall, the study shows that reconstruction-error-based anomaly detection can support the transition from experience-based lifetime test evaluation toward data-driven predictive quality assessment. With broader validation data, improved threshold calibration, and additional sensor-based features, this framework can be further developed into a more reliable decision support system for lifetime testing.

Acknowledgement

This study was supported by Eskisehir Technical University Scientific Research Projects Commission under grants no: 26LÖP047.

References

- Ahmad, S., Styp-Rekowski, K., Nedelkoski, S., and Kao, O., Autoencoder-based condition monitoring and anomaly detection method for rotating machines, *IEEE International Conference on Big Data (Big Data)*, pp. 4093-4102, 2020.
- Aslam, M. M., Tufail, A., De Silva, L. C., Apong, R. M., and Namoun, A., An improved autoencoder-based approach for anomaly detection in industrial control systems, *Systems Science & Control Engineering*, 2024.
- Carbery, C., Woods, R., and Marshall, A. H., A Bayesian network based learning system for modelling faults in large-scale manufacturing, *IEEE International Conference on Industrial Technology (ICIT)*, pp. 1357–1362, 2018.
- Chicco, D. and Jurman, G., The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC Genomics*, vol. 21, 2020.

- Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., and Barbosa, J., Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges, *Computers in Industry*, vol. 123, 2020.
- Givnan, S., Chalmers, C., Fergus, P., Ortega-Martorell, S., and Whalley, T., Anomaly Detection Using Autoencoder Reconstruction upon Industrial Motors, *Sensors*, 2022.
- Kausik, A. K., Rashid, A. B., Baki, R. F., and Maktum, M. J., Machine learning algorithms for manufacturing quality assurance: A systematic review of performance metrics and applications, *Array*, 2025.
- Krawczyk, B., Learning from imbalanced data: open challenges and future directions, *Progress in Artificial Intelligence*, pp. 221–232, 2016.
- Lee, Y., Park, C., Kim, N., Ahn, J., and Jeong, J., LSTM-Autoencoder Based Anomaly Detection Using Vibration Data of Wind Turbines, *Sensors*, 2024.
- Lee, J., Qiu, H., Yu, G., and Lin, J., Bearing Data Set, *IMS, University of Cincinnati; NASA Ames Research Center, Moffett Field, CA, USA*, 2007.
- Li, P., Pei, Y., and Li, J., A comprehensive survey on design and application of autoencoder in deep learning, *Applied Soft Computing*, 2023.
- Li, Z., Wang, Y., and Wang, K. S., Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: Industry 4.0 scenario, *Advances in Manufacturing*, pp. 377–387, 2017.
- Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., and Shroff, G., LSTM-based encoder-decoder for multi-sensor anomaly detection, *arXiv preprint*, 2016.
- Tercan, H. and Meisen, T., Machine learning and deep learning based predictive quality in manufacturing: a systematic review, *Journal of Intelligent Manufacturing*, pp. 1879-1905, 2022.
- Tuli, S., Casale, G., and Jennings, N. R., TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data, *Proceedings of the VLDB Endowment*, 2022.
- Wang, F., Jiang, Y., Zhang, R., Wei, A., Xie, J., and Pang, X., A Survey of Deep Anomaly Detection in Multivariate Time Series: Taxonomy, Applications, and Directions, *Sensors*, 2025.
- Zhang, S., Zhai, J., Chen, J., and He, Q., Autoencoder and its various variants, *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 415–419, 2018.
- Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., and Gao, R. X., Deep learning and its applications to machine health monitoring, *Mechanical Systems and Signal Processing*, pp. 213–237, 2019.

Biographies

Yigit Elarslan is an undergraduate student in Industrial Engineering at Eskişehir Technical University, Türkiye. His work has focused on the development of computer vision, navigation, and guidance systems for UAV platforms. He has also contributed to artificial intelligence solutions for various data-driven systems.

Ceren Eda Cicek is a senior Industrial Engineering student at Eskişehir Technical University with interests in operations research, optimization, and data-driven decision support systems. Currently, she is developing an AI-based Predictive Quality model for anomaly detection to support earlier and more reliable decision-making processes.

Efe Ozkaynak is a senior Industrial Engineering student at Eskişehir Technical University, focusing his academic studies on optimization, simulation, and data science. With an interest in the practical applications of evolutionary algorithms, he has also worked on an exam scheduling system utilizing a Genetic Algorithm to address multiple constraints.

Gulcin Dinc Yalcin is an Associate Professor of Industrial Engineering at Eskişehir Technical University, Turkey. She obtained her doctorate degree at Anadolu University, Turkey, and subsequently held the position of post-doctoral researcher at the Department of Industrial and System Engineering at Lehigh University, Bethlehem, Pennsylvania, USA. Her research interests lie in the fields of nonsmoothed optimization and multi-objective optimization, with a focus on their applications in various areas, including machine learning, logistics problems particularly in the context of vehicle routing, and location problems.

Eren Guvenilir is a Lead Specialist at Haier Europe's tumble dryer factory in Eskişehir, where he leads the Research & Development Laboratory. At the core of his professional role lies the mission of bringing tumble dryers up to international standards. To this end, he leads a dedicated team of engineers and technicians, overseeing the development and design of components and software to achieve target performance levels and energy classifications. He is responsible for crafting comprehensive test plans for R&D projects and serves as a pioneer of innovative initiatives. Alongside his professional responsibilities, he is currently completing his PhD in Physics, in Eskişehir Osmangazi University, specializing in condensed matter physics, with a particular focus on industry-friendly carbon-based supercapacitors and their practical applications.